

INVESTIGATION INTO THE FEATURES OF WRITTEN DISCOURSE AT LEVELS B2 AND C1 OF THE CEFR

Daniel Waller

This is a digitised version of a dissertation submitted to the University of Bedfordshire.

It is available to view only.

This item is subject to copyright.

**INVESTIGATION INTO THE FEATURES OF WRITTEN
DISCOURSE AT LEVELS B2 AND C1 OF THE CEFR**

DANIEL WALLER

Ph.D

2015

UNIVERSITY OF BEDFORDSHIRE

**INVESTIGATION INTO THE FEATURES OF WRITTEN
DISCOURSE AT LEVELS B2 AND C1 OF THE CEFR**

by

DANIEL WALLER

A thesis submitted to the University of Bedfordshire in partial fulfilment of
the requirements for the degree of Doctor of Philosophy

January 2015

Abstract

Validation in language testing is an ongoing process in which information is collected through investigations into the design, implementation, products and impacts of an assessment (Sireci, 2007). This includes the cognitive processes elicited from candidates by a test (Weir, 2005). This study investigated the English Speaking Board's ESOL International examinations at levels B2 and C1 of the CEFR. The study considered the role of discourse competence in successful performances through examination of cognitive phases employed by candidates and metadiscourse markers and whether the use fit with models such as the CEFR and Field (2004) and so contributed to the validation argument.

The study had two strands. The process strand of the study was largely qualitative and focussed on the cognitive processes which candidates used to compose their texts. Verbal reports were carried out with a total of twelve participants, six at each level. The product strand of the study analysed the use of metadiscourse markers in the scripts of sixty candidates in order to identify developing features of discourse competence at levels B2 and C1.

The process strand of the study identified that there were statistically significant differences in the cognitive phases employed by the participants in the study. The investigation also identified a number of differences in what B2 and C1 learners attended to while carrying out the different phases. The product strand of the study found no statistically significant differences in the use of metadiscourse markers used by candidates at the two levels, but observed differences in the way particular metadiscourse markers were employed. These differences indicate the direction for a possible larger-scale study.

Unlike previous studies into metadiscourse (Burneikaite, 2008; Plakans, 2009; Bax, Nataksuhara & Waller, forthcoming) the study controlled for task, text type and rhetorical pattern and nationality. The study suggested that discourse competence contributed to higher-level performances in writing and that the examinations under investigation elicited a wide range of cognitive phases from C1 candidates. The study also suggested that many of the CEFR's statements about the development of discourse competence at the higher levels are correct.

DECLARATION

I declare that this thesis is my own unaided work. It is being submitted for the degree of Ph.D at the University of Bedfordshire.

It has not been submitted before for any degree or examination in any other university.

Daniel Waller

22nd January 2015

Table of Contents

Abstract.....	iv
List of Tables and Illustrations	xvii
<i>Tables</i>	xvii
<i>Figures</i>	xxi
Commonly used abbreviations in the study	xxiii
Acknowledgements.....	xxiv
Chapter One: Introduction	1
1.1 Background	1
1.2 Structure of the Literature Review	4
1.3 The Structure of the Thesis	6
Chapter Two: Validity in Tests of Writing	8
2.1 Writing in a Second Language	8
2.2 Assessment.....	9
2.3 The Concept of Validity and Different Models	10
2.3.1 Sireci's model of validity	12
2.3.2 Claims of validity in writing tests.....	13
2.4 Definition of cognitive validity	15
2.4.1 Cognitive and metacognitive strategies.....	16
2.5 Issues around cognitive models of writing	18
2.5.1 The first problem: studies of cognitive processing	19
2.5.1.1 Gathercole and Baddeley (1993).....	19
2.5.1.2 Cognitive models of writing.....	23
2.5.1.3 Hayes and Flower (1980) and Hayes (1996).....	24
2.5.1.4 Bereiter and Scardamalia (1987).....	25

2.5.2 The second problem: methodology used in studies of mental processing	28
2.5.2.1 Verbal protocols.....	29
2.5.2.2 Veridicality.....	31
2.5.2.3 Reactivity.....	32
2.6 The model of cognition of writing in the study.....	34
2.7 The model proposed by Field	35
2.7.1 Macroplanning	38
2.7.2 Organisation	39
2.7.3 Microplanning	42
2.7.4 Translation (and Execution)	43
2.7.5 Monitoring	44
2.7.6 Revising	44
2.8 Studies of cognitive processing in L2.....	45
2.9 Conclusion.....	457
Chapter Three: Writing and Communicative Competence	49
3.1 Introduction.....	49
3.2 Validity in <i>a priori</i> assessment design.....	49
3.3 The concept of communicative competence	50
3.4 The origins of communicative competence	50
3.5 Models of communicative competence.....	52
3.5.1 Canale and Swain (1980)	52
3.5.2 Canale (1983)	54
3.5.3 Bachman (1990) and Bachman and Palmer (1993)	55
3.5.4 Celce-Murcia, Dornyei and Thurrell (1995)	56
3.6 Discourse competence	58

3.6.1 Definitions of discourse competence	58
3.6.2 The CEFR and discourse competence	63
3.7 The role of discourse competence in determining level in the CEFR.....	67
3.8 Discourse competence in writing criteria	69
3.9 Conclusion.....	70
Chapter Four: Metadiscourse Markers	72
4.1 Introduction.....	72
4.2 Features of discourse.....	72
4.3 Metadiscourse	74
4.3.1 The functional analysis of metadiscourse.....	76
4.3.2 Textual and interpersonal functions and language proficiency..	78
4.4 Metadiscourse markers as evidence of discourse competence ..	80
4.5 Metadiscourse schemes.....	82
4.5.1 Hyland's metadiscourse schemes.....	84
4.6 Categories of metadiscourse marker.....	87
4.6.1 Interactive metadiscourse markers	88
4.6.1.1 Transition markers.....	88
4.6.1.2 Framemarkers.....	92
4.6.1.3 Code glosses.....	93
4.6.1.4 Endophoric markers.....	93
4.6.1.5 Evidentials.....	94
4.6.2 Interactional metadiscourse markers	95
4.6.2.1 Attitude markers.....	95
4.6.2.2 Hedges and boosters.....	95
4.6.2.3 Self-mention.....	97

4.6.2.4 Engagement markers.....	98
4.7 Studies into metadiscourse and second-language learners	98
4.8 Conclusion	102
Chapter Five: The Pilot Study – Methods, Results and Implications	104
5.1 Introduction.....	104
5.1.1 The approaches used in the pilot study	105
5.2 Data collection methods for the process study pilot	108
5.2.1 Design of the pilot study	109
5.2.2 Participants.....	110
5.2.3 Design of the verbal reports	110
5.2.4 The pre-task	114
5.2.5 The essay writing task	115
5.2.6 The researcher's script.....	115
5.2.7 The writing paper and the question sheet	116
5.2.8 Field notes.....	116
5.2.9 Procedures used for data collection.....	117
5.2.10 Ethical considerations	117
5.2.11 Transcription	118
5.3 Analysis of the scripts for the process strand of the pilot study .	119
5.4 Data collection product.....	120
5.4.1 Design of the product pilot study	121
5.4.2 Corpus Size.....	124
5.4.3 Text types in the corpus.....	125
5.4.4 The source of the texts	126
5.4.5 Approval of the scripts	126

5.4.6 Transcription and coding	127
5.4.7 Ethical issues	128
5.5 Analysis of the product data	128
5.6 Results from the process pilot	129
5.7 Results from the product pilot.....	130
5.8 Conclusions and discussion	133
5.8.1 Discussion of results from the process strand of the pilot study.....	133
5.8.2 Discussion of results from the product strand of the pilot study	134
5.9 Limitations from the study and implications for the main study..	135
5.9.1 Research questions.....	136
5.9.2 Selection of the writing tasks.....	137
5.10 Limitations of the process strand of the pilot and implications for the main study.	137
5.11 Limitations of the product strand of the pilot and implications for the main study.	139
5.12 Conclusion.....	139
Chapter Six: Research Methods for the Main Study.....	141
6.1 Introduction.....	141
6.1.1 Research questions for the study	142
6.2 Approach to the research	143
6.3 Data collection: Process	144
6.3.1 Participants	144
6.3.2 Design of the verbal reports.....	145
6.3.3 The Pre-Task	147
6.3.4 The essay writing task.....	148

6.3.5 The researcher's script	148
6.3.6 The writing paper and question sheet	150
6.3.7 Field notes	150
6.3.8 Procedures used for data collection	151
6.3.9 Ethical considerations	153
6.3.10 Transcription	154
6.4 Data analysis: Process	155
6.4.1 Coding.....	156
6.4.2 Stage 1: Initial coding.....	158
6.4.3 Coding into NVivo.....	165
6.4.4 Field's cognitive phases.....	165
6.5 Data collection: Product.....	169
6.5.1 The case for quantitative data	169
6.5.2 The design of the study.....	170
6.5.3 Corpus size.....	170
6.5.4 Text types in the corpus.....	171
6.5.5 Source of the texts.....	174
6.5.6 Approval of the texts.....	174
6.5.7 Transcription and coding.....	176
6.5.8 Independent verification of script levels.....	176
6.5.9 Ethical issues	179
6.6 Data analysis: Product.....	179
6.6.1 Coding of metadiscourse markers.....	180
6.6.2 Descriptive analysis of the data	182
6.6.3 Mann-Whitney U Test for differences between B2 and C1.....	182

6.7 Conclusion.....	186
Chapter Seven: Results from Process Strand of the Study.....	187
7.1 Introduction.....	187
7.2 Recap on transcription & NVivo analysis.....	188
7.3 Descriptive statistics based on quantitative analysis of transcripts	188
7.3.1 Data from the verbal reports.....	188
7.3.2 Post-interview data.....	192
7.4 Inferential statistics based on quantitative analysis	196
7.5 Qualitative comments by category	200
7.5.1 Task Assessment.....	200
7.5.2 Generating Content.....	201
7.5.3 Consider Audience and Word Count.....	205
7.5.4 Text-Level Organisation.....	206
7.5.5 Immediate Planning.....	208
7.5.6 Linking Paragraphs and Summarising Content.....	210
7.5.7 Searching for Lexis.....	212
7.5.8 Monitoring (Unspecified, Content and Language).....	214
7.5.9 Revising.....	217
7.6 Conclusion for results from the investigation from the process strand of the study	218
Chapter Eight: Results from the Product Strand of the Study ...	221
8.1 Introduction.....	221
8.2 Descriptive Statistics	221
8.2.1 Recap on the text-selection process.....	221
8.2.2 Types, tokens and means.....	223

8.3 Inferential statistics	2266
8.4 Research Question Three	2288
8.4.1 Hypothesis One.....	228
8.5 Research Question Four	2288
8.5.1 Hypothesis Two.....	228
8.5.2 Hypothesis Three.....	229
8.5.3 Hypothesis Four.....	230
8.6 Research Question Five	2333
8.6.1 Hypothesis Five.....	233
8.6.1.1 Code Glosses (interactive).....	234
8.6.1.2 Sequencing (interactive).....	235
8.6.1.3 Transition Markers (interactive).....	237
8.6.1.4 Boosters (interactional).....	239
8.6.1.5 Self-mention (interactional).....	241
8.6.1.6 Hedges (interactional).....	242
8.7 Conclusions.....	2433
Chapter Nine: Discussion	2455
9.1 Introduction.....	2455
9.2 Research Questions One and Two:	2466
9.2.1.1 Macroplanning (Task assessment, generating content, consider audience and word count).....	247
9.2.1.2 Macroplanning and the models.....	249
9.2.2.1 Organisation.....	254
9.2.2.2 Organisation and the models.....	256
9.2.3.1 Microplanning (Immediate Planning, Summarising Content, Linking Paragraphs).....	258

9.2.3.2 Microplanning (Immediate Planning, Summarising Content, Linking Paragraphs).....	259
9.2.4.1 Translation (Searching for Lexis).....	262
9.2.4.2 Translation (Searching for Lexis).....	262
9.2.5.1 Monitoring (Monitoring for Unspecified Purposes, Monitoring Content, Monitoring Language).....	264
9.2.5.2 Monitoring (Monitoring for Unspecified Purposes, Monitoring Content, Monitoring Language) and the models.....	265
9.2.6.1 Revising.....	266
9.2.6.2 Revising and the models.....	267
9.3 Research Question Three:	2688
9.4 Research Question Four:	26969
9.6 Research Question Five:	2755
9.7 Conclusion.....	278
Chapter Ten: Limitations and Implications of the Study for Language Testing, Teaching and Further Research	2800
10.1 Introduction.....	2800
10.2 Conclusion: The aims of the study	2800
10.3 Limitations of the study.....	2802
10.4 Implications for language testing, the ESB ESOL International English Language Examinations, teaching and further research...	2865
References	2888

Appendix One:	Pilot verbal reporting script & prompts.....	303
Appendix Two:	Pilot written report script & prompts.....	319
Appendix Three:	Pilot process study written reports.....	321
Appendix Four:	Pilot process study verbal reports.....	342
Appendix Five:	Main study research tools (verbal reports).....	363
Appendix Six:	Information sheets and consent forms (English and Greek).....	367
Appendix Seven:	Main study verbal reports, interviews & Field Notes.....	375
Appendix Eight:	Tasks for script raters (product study).....	551
Appendix Nine:	Data from the verbal reports.....	555
Appendix Ten:	Mapping of categories onto Field's cognitive phases.....	557
Appendix Eleven:	Descriptive statistics for metadiscourse markers by category.....	560

List of Tables and Illustrations

Tables

Table 1.1	CEFR Levels	1
Table 2.1	Cognitive processing framework	36
Table 3.1	Components of the three communicative language competences as set out by the CEFR	65
Table 4.1	Hyland's categories of metadiscourse markers	85
Table 4.2	Hyland's representation of internal and external roles of transition markers	90
Table 5.1	Essay tasks used in the pilot	124
Table 5.2	Scripts collected for analysis	126
Table 5.3	Instances of cognitive processes elicited from participants	130
Table 5.4	Overview of texts analysed	131
Table 5.5	Discourse markers by level (average % of text) for Task A (C1 level task)	132
Table 5.6	Discourse markers by level (average % of text) for Task B (B2 level task)	133
Table 6.1	Overview of participants in the verbal report study	145
Table 6.2	Categories identified from the initial coding	159
Table 6.3	Researcher and Rater Trial of the Coding Categories	161
Table 6.4	Agreement by category	162
Table 6.5	Cognitive phases in the final rating scheme	164
Table 6.6	Mapping of categories onto Field's cognitive phases	166
Table 6.7	FACETS Raters Measurement Report (arranged by mN)	178

Table 6.7	Lexical exponents added to Hyland's 2005 categories	180
Table 6.8	Summary of Research Questions and Hypotheses for the product analysis of metadiscourse markers	185
Table 7.1	Number of nodes coded in B2 and C1 transcripts	189
Table 7.2	Mean number of interventions by participant	191
Table 7.3	Independent samples t-test for all categories used in the verbal reports (N=12)	197
Table 7.4	Independent samples t-test for Field's cognitive phases used in the verbal reports (N=12)	199
Table 7.5	Means for Generating Content across writing time (N=12)	202
Table 7.6	Means for Text-Level Organisation across writing time (N=12)	206
Table 7.7	Instances of note-taking prior to writing by participants	208
Table 7.8	Means for Immediate Planning across writing time (N=12)	209
Table 7.9	Frequency of lexis used by participants in the category Immediate Planning	209
Table 7.10	Most frequent words in the Summarising Content category	211
Table 7.11	Frequency of words used for Searching for Lexis	214
Table 7.12	Means for Monitoring Content across writing time (N=12)	215
Table 7.13	Means for Monitoring Language across writing time (N=12)	216

Table 7.14	Means for Revising across writing time (N=12)	217
Table 8.1	Interactive Metadiscourse categories based on Hyland 2005 and additional items	222
Table 8.2	Interactional Metadiscourse categories based on Hyland 2005 and additional items	223
Table 8.3	Descriptive statistics for types and tokens	224
Table 8.4	Mean of metadiscourse markers as percentage of texts	224
Table 8.5	Comparison of interactional and interactive markers as means and percentage of text	226
Table 8.6	Significance from one-sample Kolmogorov- Smirnov test for metadiscourse markers by category	227
Table 8.7	Mann-Whitney U test on overall totals for interactive metadiscourse markers compared by level (B2 and C1)	229
Table 8.8	Mann-Whitney U test on overall totals for interactional metadiscourse markers compared by level (B2 and C1)	229
Table 8.9	Results of Mann-Whitney U test for statistical significance of interactive metadiscourse markers by category	230
Table 8.10	Results of Mann-Whitney test for statistical significance of interactional metadiscourse markers by category	231
Table 8.11	Descriptive statistics for the number of different individual metadiscourse marker used by category	232
Table 8.12	Mann-Whitney U test for range of markers used for Boosters Self-mention and Hedges.	233

Table 8.13	Comparison of most frequently used Sequencing items	235
Table 8.14	Use of different Transition Markers according to mean	238
Table 8.15	Examples of Boosters used by B2 and C1 candidates with English Profile level of exponents	240
Table 8.16	Words occurring to the left and right of Booster 'of course' in B2 and C1 candidate scripts	240
Table 8.17	Mann-Whitney U test results by individual lexical exponents for Self-mention	241
Table 8.18	Comparison of Hedges used by B2 and C1 candidates based on mean	242
Table 9.1	Word frequency for all participants from the Monitoring phase node (N=12)	265
Table 9.2	Word frequency scores for the Revision phase node	267

Figures

Figure 2.1	Cognitive components of writing skill	37
Figure 3.1	Bachman's components of organisational competence	56
Figure 3.2	Schematic representation of communicative competence	57
Figure 4.1	Features of discourse competence in the CEFR	73
Figure 4.2	A process genre approach to teaching writing.	76
Figure 4.3	External (propositional/ideational) and internal (metadiscourse/interpersonal) roles of transition markers	90
Figure 5.1	Training task used with participants in the process pilot study	114
Figure 6.1	Procedures for data collection in verbal reports	151
Figure 6.2	Excel worksheet for analysis of metadiscourse markers	182
Figure 7.1	B2 and C1 responses to the post-interview question 'How did you prepare to write the essay?'	193
Figure 7.2	What did you think about after you had completed the piece of writing?	194
Figure 7.3	When you produce an essay like this in an exam, who do you think you are writing for?	195
Figure 8.1	Comparison of metadiscourse use by mean	225
Figure 8.2	Comparison of metadiscourse markers used as percentage of texts	225
Figure 8.3	Comparison of 4 types of Code Gloss	235
Figure 8.4	Concordance lines for 'last but not least'	236
Figure 8.5	Concordance lines for 'although' from B2	

Commonly used abbreviations in the study

CEFR	Common European Framework of Reference for Languages
ESB	English Speaking Board
L1	Language 1 (i.e. an individual's native language)
L2	Language 2 (i.e. the target language that is being learnt – in this case English)
UCLan	University of Central Lancashire

Acknowledgements

I would like to thank all of the following for their advice, help and support in the preparation of this piece of work.

First of all my supervisors Tony Green and Stephen Bax and all of those at CRELLA who have provided suggestions, support and constructive criticism throughout the project.

All my colleagues at UCLan who have helped with this project and in particular Chris Aldred, Judith Broadbridge, Nick Gregson, Nicola Halenko, Tania Horak, Chris Jones and Josie Leonard. I would also like to thank the Deans of the School of Language, Literature and International Studies at UCLan who have supported this work, particularly Isabel Donnelly.

Thank you to everyone at English Speaking Board for their support throughout the project.

Huge thanks to Europalso for their help in Athens and in particular to Anna Bouldoumi, Jane Collins, Emmanouella Christodoulou and George Zikopoulos. I would also like to thank the all learners who participated in the study for their time and efforts.

Finally, I would like to thank my wife Imren and son Teoman for the time and space given to me in which to work on this project.

Chapter One: Introduction

1.1 Background

This study was initiated as part of the development of the English Speaking Board (ESB) ESOL International Examinations by the University of Central Lancashire. The ESB ESOL International Examinations are administered in Greece to around 20,000 candidates annually at levels B1, B2, C1 and C2 of the Common European Framework of Reference for Languages (CEFR). The CEFR (Council of Europe, 2001) is a document which describes levels of language attainment. There are nine levels in the CEFR's scheme (see Table 1.1 below) which can be grouped under three main descriptions (Basic User, Independent User and Proficient User).

Table 1.1 CEFR Levels

Basic User	A1	Breakthrough
	A2	Waystage
	A2+	
Independent User	B1	Threshold
	B1+	
	B2	Vantage
	B2+	
Proficient User	C1	Effective Operational Proficiency
	C2	Mastery

The candidates for the ESB exams are mainly students studying at 'frontistiria', the private language schools which learners attend in Greece after formal schooling hours. The age of students taking the ESB ESOL

International Examinations varies but students taking the B1 tend to be the youngest, aged between 13 – 14, while the candidates taking B2 will be 14 – 15, C1 15 – 16 and C2 candidates are likely to be aged 16 and over. The express intention of many of these students is to be ‘done’ with English by the time they enter university. By this they mean that they will have passed a C2 level English language qualification. The C2 level has an additional attraction for learners in that it is regarded by the Greek Ministry of Education as being the level at which someone can be considered qualified to teach English and up until recently a C2 qualification was sufficient to allow someone to open a frontistiria of their own¹.

The ESB ESOL International qualifications were developed by the University of Central Lancashire (UCLan), which still produces the papers and conducts quality assurance and research on behalf of ESB. The current format of the examination, including the style of essay questions has been in use since 2006, following two years of trialling and feedback. However, as the ten-year anniversary of the examinations approaches there is a desire on the part of all involved to develop the examinations further and to revisit the theoretical underpinning of the formats and criteria used. This study is intended as one of a set of investigations which will assist in the revision of the ESB ESOL International Examinations.

The ESB ESOL International Examinations assess the skills of listening, reading, writing and speaking and are intended to be a test of communicative English language proficiency. The decision was taken to focus on the writing section of the examinations in this study due to an interest in evaluating the extent to which the essay task used elicited a sufficient range of competencies from candidates for valid assessment.

¹ This requirement has now been amended with an individual requiring a recognised TESOL qualification in addition to a pass at the C2 level.

Additionally, the researcher wished to investigate whether candidate writing produced under timed examination conditions displayed the discourse features which the CEFR states are the identifying features of higher-level performances in writing.

Not all candidates in the B2 and C1 ESB ESOL International Examinations choose to produce an essay because a choice of written task types is offered. Despite this, essays were chosen as the genre for investigation in this study because this type of task is frequently set in English language assessments at the (CEFR) independent (B) and proficient (C) user levels (Council of Europe, 2001). The B2 and C1 levels were specifically chosen as the levels for the investigation because they are the levels at which the entry criteria onto university programmes of study in the UK are often set (IELTS, 2012).

Discourse is emphasised in the CEFR as a key feature of performance which characterises the B2 and C1 levels. The CEFR states that at the B2+ level and beyond there is “a new focus on discourse skills” (Council of Europe, 2001, p. 35) which implies that for candidates writing at these levels there should be an increased awareness of macro-features such as genre, audience and text purpose and of how micro-features such as organisation and discourse markers contribute to the target genre. This study investigates the texts produced by candidates for evidence of such discourse features.

The second area of investigation in this study is the question of how second language writers go about composing their essays under timed examination conditions. If, as stated above, writers at the B2+ and C1 levels do have more awareness of discourse, then presumably consideration of genre, audience and purpose will increasingly be a feature of the composition process. In addition to this there is the question of whether the actual tasks set in the ESB ESOL International

Examinations at the B2 and C1 levels allow candidates sufficient scope to demonstrate these processes and thereby display the increased skill in discourse which the CEFR asserts is such a prominent feature at these levels. Therefore, the study will also investigate the process of composition used by candidates at the B2 and C1 levels in the ESB ESOL International Examinations. Since the cognitive processes used by candidates are internal and cannot be directly observed, evidence to imply these processes will be sought from observation of the candidates through the use of verbal reports. The aim of this strand of the study is to search for evidence of cognitive validity; that is whether the test “represents the cognitive processing involved in writing contexts beyond the test itself, i.e. in performing the task in real life” (Shaw and Weir, 2007, p. 34). The concept of cognitive validity is explored further in section 2.4.

As set out above, this thesis has two main aims which the research questions will seek to investigate and answer through the examination of the processes and products elicited from candidates by the essay tasks. These aims are:

1. To what extent is cognitive validity demonstrated in the cognitive phases that candidates carry out while producing scripts in the English Speaking Board ESOL International Examinations?
2. What is the role of discourse competence in deciding whether a script is classified as being level B2 and C1 of the Common European Framework of Reference for Languages (CEFR) in candidate scripts from the ESB ESOL International Examinations?

1.2 Structure of the Literature Review

The literature review is divided into three chapters. Chapter Two will begin by discussing the issue of validity as a central tenet of language assessment and the need for assessment developers to demonstrate evidence that the task types used in assessments are appropriate to the constructs being elicited and the purposes to which the results of the

examination will be put. Central to the issue of validity is whether tasks are able to elicit the actual mental processes which candidates would be expected to use in the real-world beyond the test. This dimension of validity has become of increasing interest in the area of investigating validity in writing (Kellogg, 1999; Field, 2004; Shaw & Weir, 2007; Barkaoui, 2011). Demonstrating that an assessment displays cognitive validity is a crucial piece of the evidence-based argument for that assessment's validity. However, the methodology used by researchers to investigate cognitive processing during writing has been controversial in the past, so Chapter Two will also consider the suitability of verbal reports as a research tool. In summary, Chapter Two aims to establish a theoretical framework for investigating cognitive validity in candidate composition processes.

Chapter Three will then proceed to discuss the concept of communicative competence and how models of competence contribute to an examination's validity. In order to assess a candidate's ability in writing, there must be a theoretical construct defining the abilities that contribute to good writing. This chapter will explore how models of communicative competence and the CEFR approach writing as a skill and the role of discourse competence in particular in characterising the higher levels of proficiency. Discourse competence as a concept is explored in 3.6 but it is defined by Celce-Murcia, Dornyei and Thurrell (1995, p. 13) as "where the bottom up lexico-grammatical microlevel intersects with the top-down signals of the macrolevel of communicative intent and sociocultural context to express attitudes and messages, and to create texts". The CEFR suggests that discourse competence is a key element in successful writing, especially at the higher levels, and predicts that successful writers will pay more heed to issues of discourse when they are producing texts. Through this argument, this chapter will connect the concept of discourse competence to the concept of validity set out in Chapter Two in order to

support the chosen methods of analysis to be used in the investigation of the products.

Chapter Four will go on to examine how the use of metadiscourse markers could provide evidence of a candidate's developing discourse competence in writing. The term metadiscourse marker will be defined and different typologies of metadiscourse schemes will be examined along with some of the issues around the use of metadiscourse markers as evidence of discourse competence.

1.3 The Structure of the Thesis

As set out above, the literature review will identify the main issues in the current literature in the areas covered in this study as well as the gaps which this project aims to fill. The literature review will be concluded by the research questions for the study which are:

1. What cognitive phases do candidates at levels B2 and C1 appear to employ when composing timed essay tasks?
2. To what extent does the timed essay question format used in the ESB ESOL International Examinations elicit the cognitive processing that models predict at levels B2 and C1?
3. Is there a difference in the quantity of metadiscourse markers used by candidates of the ESB ESOL International Examinations at levels B2 and C1 of the CEFR?
4. What are the functions of the metadiscourse markers used by candidates at level B2 and C1?
5. To what extent do the metadiscourse functions employed by candidates at levels B2 and C1 match the predictions made by the CEFR regarding the development of discourse competence in learners at these levels?

The literature review will be followed by Chapter Five which will set out the research methods used in the pilot study, the results of this study and the lessons learned for the main study.

Chapter Six builds on the methods used in the pilot study and details the research methods for the main study investigation into cognitive phases and products. This chapter reflects the two strands of investigation in the study. It begins by setting out the research and analysis methods for the process strand of the study and then does the same for the product strand of the study.

This division between the process and the product investigations of the role of discourse competence continues in the following two chapters. Chapter Seven provides the results from the process strand and Chapter Eight the results for the investigation of the products.

Chapter Nine brings **the** two strands of the study back together. In this chapter, the research questions are used to structure the discussion of the results and the conclusions reached regarding the aims of the study.

Chapter Ten explores the limitations of the study as well as considering implications and areas for further research.

Chapter Two: Validity in Tests of Writing

The chapter opens with a very brief contextualisation of writing in language testing before moving on to consider the notion of validity in language assessment. The focus will be on the role of mental processing in writing tasks and how assessors take cognitive validity into account in the design of such tests. Models of cognitive processing in writing will then be discussed and evaluated in order to identify a workable model which can be applied later in the study.

2.1 Writing in a Second Language

Writing in a second language is one of the four key skills which learners of a language are usually expected to acquire. In traditional approaches to second language learning such as the Grammar Translation Method, which dominated language teaching for centuries (Richards & Rodgers, 2001, p. 3-7), writing was considered alongside reading to be the essential skill. The assessment of written language in the Grammar Translation Method would consist of the mediation of prestigious literary forms such as translations from the classics or the production of essays. Learners of English were likewise expected to demonstrate the ability to produce these prestigious forms in the target language. Milanovic and Weir (in Shaw & Weir, 2007, p. xii) note that as far back as the Certificate of Proficiency in English examination in 1913, candidates were expected to produce an essay composition in two hours.

Since that time, writing in a second language as a skill has decreased and increased in importance depending on the prevailing teaching methodology of the day. The skill was downgraded in some methods such as the Audiolingual Method before being reinstated as an essential skill in the current mainstream of English Language Teaching (ELT) which

employs a broadly communicative paradigm (Thompson, 1996). As such, the skill is now an expected component of proficiency examinations and is usually assessed by requiring candidates to produce particular text types under timed and controlled conditions. For those setting the examinations there are a range of issues that must be considered; what type of text should the candidates be asked to produce?; how long should it be?; how much input should be allowed (e.g. can candidates use dictionaries?); what is the candidate expected to produce by way of response to the task? What characteristics should the piece of writing demonstrate in order for it to be considered a 'pass'? The task bears a large part of the responsibility for eliciting the written sample to be assessed but the criteria against which the candidate's work is measured are of equal if not greater importance in the determining the success of the piece of writing.

2.2 Assessment

At this point the purpose of assessments will be considered before moving onto the discussion of the importance of validity in language testing. It is also important to clarify some of the terms being used in this study. In the literature, *assessment* is usually used as a superordinate, covering a wide range of activities including written and practical examinations, coursework and even quizzes (Brown, 2004, p. 4). For the purposes of this piece of work, I will use the terms *assessment* and *test* and *testing* interchangeably. However, the term *examination* will be used with reference to timed written assessments carried out under controlled examination conditions. Brown (2004: p. 4) defines test (or assessment) as being "a method of measuring a person's ability, knowledge, or performance in a given domain". A key part of this definition is the notion of measurement; one cannot measure without knowing what the unit of measurement is. Measurement is usually understood as dealing with specific quantities; time, distance, temperature etc., however, in language testing the other terms from Brown's definition are the objects which are being measured: performance, ability or knowledge. In short, in order to

determine whether a piece of writing is successful there must be a concept of what constitutes a successful piece of writing (appropriate to the level) and what qualities must be considered. Those setting examinations, in order to produce meaningful assessments, must be able to turn to some model of successful language use in order to describe what features are to be elicited and marked.

Since its publication in the year 2001, the CEFR has become one of the most widely used models² of language use. Examining organisations work hard to demonstrate that their assessments are effectively linked to the CEFR through the use of empirical evidence. Part of this continuous process of demonstrating that a test is aligned is showing that the tasks used for the assessment of candidates are valid and accurately represent the target domain use for which the test is intended and that the responses elicited from candidates are useful in determining the relevant abilities and/or knowledge of those being assessed.

2.3 The Concept of Validity and Different Models

At this stage the discussion will now turn to the issue of validity since the current project aims to investigate whether the writing tasks in the examination under investigation can provide evidence of the development of discourse competence in test-takers. This discussion will begin by defining the term validity and considering the concepts which the term incorporates.

Validity is often regarded alongside reliability as being one of the cornerstones of effective and ethical testing. While reliability is concerned with the consistency of test results or “the extent to which the same rank order of candidates is replicated in two separate (real or simulated) administrations of the same assessment” (Council of Europe, 2001,

² For a discussion on the issue of the CEFR as a *model* rather than a *framework* see the discussion in Chapter Three, 3.6.2.

p.177), validity centres on the trustworthiness of the data extracted by a test and the application of this data to decisions made about the candidates. The CEFR states that an assessment has validity if “it can be demonstrated that what is actually assessed (the construct) is what, in the context concerned, *should* be assessed, and that the information gained is an accurate representation of the proficiency of the candidates concerned.” (Council of Europe, 2001, p.177. Italics in original).

Bachman (1990, p. 236-238) highlights that although different types of validity are often discussed, establishing the validity of an assessment is a process in which data is collected in response to different questions. In essence, ‘validation’, rather than validity, is an argument which is built on an on-going and continuous basis. This model of validity has its origins in the work of Messick (1989). Messick, building on the work of Cronbach (1971), took issue with the traditional notion of validity as being made up of different components; content validity, construct validity and criterion validity and the view, as stated by Ruch (1924, p. 13) that:

“For an examination to possess validity it is necessary that the materials actually included be of prime importance; that the questions sample widely among the essentials over which complete mastery can reasonably be expected...and that proof can be brought forward that the test elements (questions) can be defended by arguments based on more than mere personal opinion.”

The view of validity which has developed from Messick’s work suggests that validity is a far more complex concept which extends beyond the test itself, thereby making it impossible to absolutely state validity for a particular assessment. Furthermore, Messick argued that while validity could not be conclusively proved, questions which collected data from a range of different sources could help to establish an argument of validity for an assessment.

2.3.1 Sireci's model of validity

Sireci (2007, p. 477) identifies four essential features of validity in what might be termed, post-Messick, the 'modern' tradition. First, validity is not a feature which is located within a test of itself. Instead the concept is related to the purpose to which the test is put. Messick (1989) argued that it was essential to state the purposes for which the test scores would be used; that is, there must be a specification of the meaning which will be applied to a test result. A test or task cannot be described as being 'valid'; only the inferences made based upon it. The second feature of validity identified by Sireci is the use of multiple sources of evidence to support the claims made based on the test. This is where the notion of the different 'facets' of validity can be described such as construct, criterion, predictive, content, face and other dimensions of validity. Further to this the outcome of a test cannot be claimed to be valid on the basis that it 'satisfies' one of these dimensions and as Green (1998, p. 22) points out, no single aspect of validity can be considered to be more important than another. Validity or validation is an on-going 'argument' which must collect evidence from all of these areas in as systematic a manner as possible in order to keep on supporting the claims made around the results of the test.

This observation by Green connects with Sireci's third feature of validity which is that only by the building of sufficient evidence to support claims of validity for a test's use for a particular purpose can its use be defended. Finally, Sireci establishes that validity is an on-going process and not a one-off evaluation. The change of emphasis in the view of validity as being iterative can be seen in Kane (2006) who proposed the shift from discussing *validity* to *validation*, thereby signalling the notion of process.

In addition to being seen as an on-going process of argument-building, various writers have brought other concepts under the umbrella of validation. Weir (2005) outlines how historically reliability was often

perceived as being distinct from validity but argues that reliability, or consistency in test results, should be regarded as a feature of what he terms scoring validity (Weir, 2005, p. 22-35). This is a logical argument as any piece of assessed writing must be marked by using criteria which define what is being looked for. Such a set of criteria must be valid in terms of representing what is to be assessed but also must be applied consistently (i.e. reliably) in order for the test to be considered fair. This example demonstrates the essential unity of validity and the role of reliability within this.

2.3.2 Claims of validity in writing tests

The discussion now turns to the question of how claims of validity can be demonstrated in tests of written English. As set out above, such claims can only be answered through a multi-faceted approach. Weir (1988) identifies two stages within which examiners can build validation evidence. The first of these areas is *a priori* construct validation. This is the specification of the examination usually during its development: the purpose of the assessment, its intended uses, the setting, the candidates, the tasks to be used and the content of the examination. *A priori* specification should also be supplemented through the use of review and piloting of the materials. *A priori* specification can then be further supported by the gathering of *a posteriori* data once the examination has been used in order to establish that the assessment is functioning as predicted. Both *a priori* and *a posteriori* investigations may make use of similar types of data in order to build a case for validation.

Weir (2005) proposed a socio-cognitive model of test specification. The name captures the two key features of validity; the mental processes carried out by the test-taker in order to carry out the task and the social purposes being tested along with the uses to which the test scores are put. In Weir's model of validity in language testing, context validity covers two main areas: the actual task, its design (e.g. including the genre, the

weighting, length and other features) and implementation (e.g. physical conditions, standardisation in terms of the administration of the assessment) as well as the linguistic requirements needed to understand the task and perform it. Cognitive validity relates to the mental processes which candidates must perform in order to carry out the set task and whether these reflect the same cognitive processes which someone carrying out the task in the 'real' world would follow. Finally, scoring validity is connected to the development of the criteria and/or rating scale and the processes around the rating process including standardisation and moderation of the raters involved. As has been discussed above, scoring validity is seen by Weir (2005) as encompassing the concept of reliability and as such necessitates that criteria used in the judgement of assessed pieces of writing assist raters to identify clear features which will assist them to accurately assess writing consistently.

All of these areas need to be specified clearly in order to set about building evidence of validity in a test of writing. In terms of *a posteriori* sources of evidence (sources of evidence that can only be gathered once a test is administered) Weir (2005) emphasises consequential validity; that is the washback and impact of the test both on the individuals but also in wider contexts including in society at large. Finally, criterion-related validity is carried out in order to demonstrate the relationship of the test with some form of external measure including how well an assessment predicts future performance in the target language domain within which the individual wishes to operate.

In order to demonstrate validity and provide data for many of the areas outlined above it is necessary for test designers to design their assessments *a priori* with reference to established theories related to the skill under examination. These theories will assist underpinning the examination and informing the decisions and judgements made by the test designers and assessors. However, as Weir's model of validity suggests it

is not sufficient simply to appeal to these theories in the design stage; data must be collected *a posteriori* to establish that the models and assumptions made are correct and defensible.

The current study aims to investigate the cognitive validity of timed essay writing by considering the processes that candidates go through when they are composing their essays. As this is an examination which is already in existence and is considering the products of the current test, it represents an *a posteriori* study. However, as stated above, a *a posteriori* investigation can involve types of data collection which would often be seen as being more relevant to a *a priori* investigation. It can also be argued that since the purpose of the study is in part to influence the future design of the examinations, there is a strong *a priori* element to this investigation. In order to do explore the cognitive validity of timed essay writing the study must also ask the question of how far the essay task allows candidates to display these cognitive processes. Factors such as task and the required cognitive processing impact on other facets of validity such as the scoring validity in that it might be asked how such features are acknowledged in marking criteria and on context validity. The impact of one facet upon others serves to underline Sirci's (2007) point about validity being in fact a single concept with multiple dimensions. However, the main focus of this study will be on the area of the cognitive validity of the assessment.

2.4 Defining cognitive validity

Bachman (1990, p. 255) discusses construct validity by citing Messick's "basic question" of "what is the nature of" the thing which is to be measured, bringing us back to the definition of testing identified by Brown in 2.2. As such, construct validity in writing must be based on two things. First of all, a theory of what constitutes good writing, which will be further explored in Chapter Three. Secondly, the models of the mental processes that a candidate is expected to use in order to carry out a writing task. Shaw and Weir (2007, p. 34) define cognitive validity as being "a measure

of how closely [a writing task] represents the cognitive processing involved in writing contexts beyond the test itself, i.e. in performing the task in real life.” It can therefore be seen that cognitive validity is an essential part of construct validity because it provides a description of at least part of what is being measured in tests of writing. Yet, clearly there are difficulties for those who wish to investigate cognitive validity in writing, mainly because the internal processes are not themselves directly observable.

As already stated, cognitive validity as a concept emerged from the earlier notion of construct validity, as outlined by Bachman (1990, p. 254-258) and the utilisation of theory-based validity, a dimension of validity which Weir (2005, p. 17) suggests was often side-lined in favour of discrete item and statistical measures. Cognitive validity requires that test developers ensure that their tasks, criteria and examination procedures are underpinned by evidenced cognitive models of the expected real-world processes involved in language use.

Shaw and Weir (2007, p. 6) propose that cognitive validity should be demonstrated *a priori* through the specification and piloting of tasks and a *posteriori* through analysis of the data, **statistical and qualitative**. As this project is investigating the case of an examination already in use, it is possible to investigate both the processes and the products in line with Shaw and Weir’s suggestions.

2.4.1 Cognitive and metacognitive strategies

A distinction often made in the literature is between the notions of cognitive and metacognitive strategies in second language learning and second language use. Canale and Swain (1980) proposed that communicative strategies were used as part of strategic competence to repair breakdowns, strategies which were based on the application of cognitive processes (Purpura, 1999). Despite this observation, Canale and

Swain's model did not expand on how these strategies fit within a model of mental processing.

Bachman (1990, p. 98 – 107), building on the work of Faerch and Kasper (1984) extended the concept of strategic competence to cover all aspects of language use, not only speaking and identified it as comprising three components: *assessment* of the communicative goal, including consideration of resources available to interlocutor and receiver and evaluation of the success of the attempted communication; *planning* and *execution*. The model was further developed by Bachman and Palmer (1994) to illustrate the interaction between linguistic knowledge, metacognitive strategies and affect in communication. However, the model was not empirically based and was limited to the application of metacognitive strategies and did not explore how cognitive strategies interacted in communication (Purpura, 1999).

It has been argued that there is a close connection between cognitive and metacognitive strategies. Phakiti (2006, p. 56) reports that metacognitive strategies, those of task assessment, planning and execution, may direct lower level cognitive strategies which are often more automated or semi-automated and involve skills such as translating, summarising, making links with previous knowledge or experience or applying lexicogrammatical rules and patterns. In short, a learner may be more consciously aware of the metacognitive strategies than the cognitive ones but also a learner may be taught when and how to use metacognitive strategies. Phakiti (2006, p.56) also makes the point that metacognitive strategies may become more automatic with increased proficiency, and thus appear to be indistinguishable from cognitive processing, despite the fact that these higher level skills are still controlling the composition process.

The development of metacognitive strategies clearly impacts on the ability of second language writers to respond effectively to a set writing task and

it can postulated that more successful language users will be more skilled in applying these strategies in their writing.

2.5 Issues around cognitive models of writing

As has already been alluded to in the previous section, the question of how cognitive validity can be demonstrated, whether by researchers into cognitive processing or by those involved in the development of language tests, is not without problems. These problems stem from the essential difficulty of accessing cognitive processes and creating a model of these and by extension ascertaining which processes are used to carry out a writing task.

First of all, there must be a model of how writers write both in general but also in producing the text type being used in the test. While there are studies which examine the mental processes which authors go through when they are producing text, many of these studies are based on writing in L1. There are a number of studies which consider how L2 writers produce a texts or task type used in an assessment and these are discussed in 2.8. A second problem is the issue of how those exploring cognitive validity access the cognitive processes and the methodology used to examine these processes.

A further issue and one identified by Weigle (2002, p. 197) concerns the written task type used in studies. In this particular study the type of task being assessed is timed essay production but that does beg the questions of where and when such tasks are indeed performed in the real world and whether such tasks can offer a wide enough range of processes to represent proficiency in writing, particularly at the higher levels of proficiency in English language.

2.5.1 The first problem: studies of cognitive processing

Different writers have proposed various models of cognitive processing in writing (Hayes and Flower, 1980; Bereiter and Scardamalia, 1987; Grabe and Kaplan, 1996; Kellogg, 1994, 1996; Hayes, 1996; Field, 2004). These models tend to demonstrate some overlaps in terms of their components in that they represent the interaction of three key factors: (i) the task and the environment of the task, (ii) the internal cognitive processes drawn on by the writer and the role of short-term memory (STM) in these and (iii) the role of the writer's knowledge, experience and faculties as held in the long-term memory (LTM). It is therefore necessary at this stage to examine some of the principles upon which these models rest, in particular the operation of cognitive process and the role of STM and LTM.

Consideration of the principles on which these models are based is essential as all of the cognitive models of the writing process rest upon research carried out into psycholinguistics and neurolinguistics and draw upon such studies in order to build models that attempt to predict the behaviour carried out by writers. In other words, the validity of the models of the writing process rest on the validity of models of cognition.

2.5.1.1 Gathercole and Baddeley (1993)

The model of memory presented by Gathercole and Baddeley (1993) is of particular relevance to this exploration of cognitive processing in writing because it is the model which is drawn upon by Kellogg (1999) and Field (2004) and it is their model which the study is using as a basis for the investigation (see 2.6 for the rationale for this).

In their model Gathercole and Baddeley follow Baddeley and Hitch's position on the short-term memory store (Eysenck, 1993) by employing the term 'working memory' rather than STM. This distinction between STM and working memory is based on the assumption in their model that the role of working memory is not only to hold information on a temporary basis but also to process it (Field, 2004, p. 326).

In the Gathercole and Baddeley model, working memory is comprised of three major components which Eysenck identifies as follows (1993, p. 71):

- A modality-free central executive, which is virtually synonymous with attention.
- An articulatory loop [often referred to as the phonological loop], which can be regarded as a verbal rehearsal system; it resembles an inner voice.
- A visuo-spatial sketch pad, which is a visual and/or spatial rehearsal system; it resembles an inner eye.

Working memory is identified by the authors as having a limited capacity. Processes are managed through the 'central executive' which coordinates the activity of working memory (Gathercole & Baddeley, 1993, p. 5) including relaying information between other parts of the system such as the two 'slave systems': the phonological loop and the sketchpad. The first of these, the phonological loop which is also sometimes referred to as the articulatory loop (Hitch, 2005, p. 315), retains verbally-coded information but this information is subject to decay over time. Information which is already phonetically coded such as spoken input can be directly held by the short-term store of working memory but other input such as written words or pictures has to be processed, that is coded into phonological form by sub-vocal rehearsal. For example, when reading there is the 'inner voice' which is often described as if often regarded as evidence of the phonological loop converting visual input into verbal form. Evidence for the phonological loop has been gathered from many sources including experiments to disrupt the process as well as from neuropsychological evidence from individuals who have suffered damage to the parts of the brain responsible for input (Gathercole & Baddeley, 1993, p. 8-17).

The second 'slave system' is the visuo-spatial scratchpad. The existence of a separate cognitive process was noted by Baddeley and Hitch (1974) who identified that subjects in experiments who were asked to carry out two verbal tasks or two visuo-spatial activities simultaneously found it more difficult than those who were required to carry out two tasks where one comprised a verbal and one comprised a visuo-spatial task. They proposed that these studies suggested that there was an extra subsystem for the storing of visuo-spatial information.

Gathercole and Baddeley draw on the work of Shallice (1988, cited in Gathercole & Baddeley, 1993, p. 5) in order to describe the way in which the action is controlled by the central executive in their model of working memory. According to Gathercole and Baddeley (p.5), the central executive is responsible for a number of functions:

- Coordinating activity within working memory.
- Coordinating the movement of information between the components of the cognitive system.
- Managing resources to be used by the phonological loop and visuo-spatial sketchpad.
- Retrieving information from the long-term memory.

Gathercole and Baddeley describe how the central executive system operates although the model is far from complete with the executive being "the most important component of working memory, the most controversial and the least understood" (Hitch, 2005, p. 323). A number of studies have illustrated this and attempted to measure the limitations of the capacity of working memory and, according to Hitch (2005, p.325), it is likely that future accounts of the central executive may be more complex and fractured. However, in the absence of such accounts, the Gathercole and Baddeley model will remain the model used in this study.

The way in which the central executive manages working memory is set out by Gathercole and Baddeley. In their description processes are started in response to environmental triggers (i.e. the external physical environment). Many forms of behaviour may be automatic and are governed by schemas which are “a memory structure that encapsulates an event” (Conway & Holmes, 2005, p. 519). For example, participating in an examination might activate a number of schema for example, checking where one’s seat is, placing bags and coats in the designated part of the room and so on. Such activities are automatic and guided by schemas and are similar to learned behaviours. For example, when driving a car environmental triggers, such as a red traffic light, are sufficient to activate the appropriate schema (i.e. the braking schema). Gathercole and Baddeley (1993, p. 6) suggest that working memory is managed by a supervisory attentional system (SAS). The SAS is employed to inhibit the schema when these are in conflict with the actual environment when normal contention scheduling is not sufficient to manage the conflict between competing schema. For example, to use the driving schema which was mentioned before, should an emergency vehicle come into sight with siren and lights on then there are two schema in competition, the *driving* schema and the *get out of the way* schema. In those situations where the situation is novel or the environment is urgent or threatening (to extend the driving example, the road is blocked and moving out of the way of the ambulance is a difficult matter), the SAS intervenes and can override processes.

The workings of the central executive are important in the discussion here because it suggests that many routine activities are automated, and it can be conjectured that writing may well be a process which uses schema to direct production. However, when the demands of the writing task are novel, a writer may respond in one of two ways. A writer may either identify that different requirements are needed and therefore begin to engage in a problem-solving approach or, if the writer is unaware that

such a different approach is needed, may attempt to apply existing schema.

A second key feature of the model of working memory proposed by Gathercole & Baddeley (1993) is the phonological loop 'slave system' which serves to store "material in phonological code which decays in time." (p.8). Gathercole and Baddeley (ibid) state that the model has a solid basis in experimental evidence and studies of neuropsychological patients. The model proposes that input, whether speech, visual or written in form, is temporarily recorded in a phonological store. Speech input can be directly stored however non-speech input (visuals or writing) is required to go through a subvocal rehearsal, effectively being coded into phonological form; the 'inner voice' that people often cite when reading or writing would be an example of this.

The strong evidence base for the model proposed by Baddeley and Hitch (1974) and Gathercole and Baddeley (1993) from experiments (see Hitch 2005, p.317-322) and neuropsychology and its components suggests that the foundation of working memory is a valid basis from which to explore the writing process. The model is not without criticisms though and understanding of the workings of the central executive, as mentioned above, are seen to be problematic and incomplete. However, the prevalence that the model still enjoys together with its evidence base and the lack of an alternative model means that this model of working memory remains viable for the purposes of this study.

2.5.1.2 *Cognitive models of writing*

In order to investigate the cognitive processes of candidates engaged in writing tasks a method must be chosen with the ability to access some of what candidates have attended to during writing and which can assist in the construction of accurate and predictive models for the writing process.

2.5.1.3 *Hayes and Flower (1980) and Hayes (1996)*

As stated at the beginning of this section, a number of researchers have investigated the cognitive processing of writers and proposed cognitive models for how writing is carried out. The chosen method frequently used in such investigations is the verbal report, a method vigorously supported by Ericsson and Simon (1980). Cognitive models of writing such as those proposed by Hayes and Flower (1980) and Hayes (1996) sought to describe the writing process in terms of the use of working memory and schema. However, by focussing on the internal processes such models of writing often faced criticism centred on the issue of the writer being portrayed as an isolated individual (Nystrand, 1982). These models were seen as neglecting the environmental and social impacts on the writer. The later version of the model put forward by Hayes (1996) added the influence of social environment on the production of writing. Despite criticisms of the Hayes and Flower models, their observations that the internal cognitive processes are recursive and not linear in nature and that these processes occur throughout the act of writing have had a major impact on subsequent models.

Other issues have been raised in relation to the Hayes and Flower (1980) and Hayes (1996) models. Weigle (2002, p. 28) states that the models fail to provide sufficient detail of how situational variables influence writing or of the role of linguistic knowledge. More importantly for those interested in the cognitive processes of writing, there is no indication as to how the different components interact during the writing process (Shaw & Weir, 2007, p. 35). Grabe and Kaplan (1996, p. 92) note a further problem in that the Hayes and Flower and Hayes models do not allow for different approaches to writing and crucially for a distinction between the approaches taken by writers with higher or lower levels of competence in the skill.

A further point to make is that the models of writing put forward by Hayes and Flower (1980) and Hayes (1996) were models of writing in the first

language. Issues around levels of writer competence and the lack of detail over the role of linguistic competence render the models particularly problematic for the discussion of second language writers where both of these issues become extremely important. Second language writers invariably will have differing levels of linguistic competence but also are likely to have very different experience of text production depending on their cultural background. Hinds (1987) proposed that some cultures have a default position of being 'writer responsible', in that it is the author who is responsible for ensuring the communicative purpose of the text is realised while other cultures are 'reader responsible' and the task of interpretation is upon the reader. While Hinds' concept of reader/writer-responsible cultures remains controversial and perhaps oversimplifies issues such as discourse, a number of studies have identified that there are variations between cultures in the way communicative acts in writing are carried out within different text types (Vergaro, 2005 ; Qi & Liu, 2007). So while a candidate in an exam may be considered an expert writer of one particular text type in their own culture, this may not translate into expertise in a second language culture. Swales' (1990) experiences of attempting to write in an unfamiliar genre illustrates the principle that for many types of discourse there may be a period of 'apprenticeship' during which the writer moves from being a novice through to being an expert.

2.5.1.4 Bereiter and Scardamalia (1987)

Bereiter and Scardamalia (1987) put forward two models which do attempt to address the issue of varying expertise in writing. The two models represent writing as *knowledge telling* or *knowledge transforming*. *Knowledge telling* is the process carried out by less adept writers and is largely concerned with the generating of content and its translation into written form. According to Bereiter and Scardamalia, writers who are knowledge-telling are less likely to be concerned with issues of *how* to write than *what* to write. The model is often used to describe the composition processes of young native-speaker writers and the typical

behaviour of interpreting a task as requiring them to 'tell' everything that they can recall on a particular issue without necessarily considering an overall line of argument or engaging with rhetorical problems such as the logical sequencing of information in the text. Those approaching writing from a knowledge-telling perspective will rely on immediately available content, topic and genre. In the case of genre this will often mean relying on tried-and-tested patterns without necessarily considering the suitability of the pattern for the task at hand. By contrast writers who employ knowledge transforming will view writing as a series of rhetorical problems at each stage of the writing process, whether it be selecting content from that generated, taking into account how to structure the information available, considering features such as audience awareness and the likely status of information (i.e. familiar or unfamiliar to the intended reader), and so on.

There are other important features in the models proposed by Bereiter and Scardamalia. Firstly, a writer who is capable of knowledge transforming will not always need to do so; some tasks, particularly if predictable and routine can be carried out using knowledge-telling processes, particularly if the writer has experience of producing the particular text type previously and is very familiar with carrying out the functions required. An example of situations where a good writer may not need to use knowledge transforming processes might be to produce personal narratives or anecdotal writing. A second principle is that being able to knowledge transform in one genre does not automatically transfer to a different one (Grabe & Kaplan, 1996, p. 125). A writer will have to deal adequately with the issues of the rhetorical form of the new genre and 'solve' the problems of managing the information and requirements of the new genre.

While the models of knowledge telling and knowledge transforming help to account for the differences in expert and non-expert writing, they nevertheless face difficulties when applied to second language writing.

First of all, the models are based on native-speaker performances and are more applicable to the development of advanced writing skills in children and young adults. However, it is not unreasonable to argue that a second language writer who is judged to be proficient in producing certain types of specialist texts, such as high-level academic work, is likely to utilise knowledge transforming processes. What is more challenging is the issue of how a writer, whether native or non-native in background, moves from knowledge telling to knowledge transforming and how does one identify the difference? If, as Grabe and Kaplan (1996, p. 124) suggest, an experienced writer may be able to draw on previous experience and complete a task through knowledge-telling processes alone, then it would be extremely difficult to determine whether a writer was knowledge telling or transforming. In a similar fashion, when it comes to the setting of writing tasks it is presumably the case that whether a task generates knowledge telling or knowledge transforming is dependent not solely upon the task itself but on the processes that the candidate brings to it. That said, it is also clear that certain types of task, such as free-writing narratives and anecdotal accounts are more likely to elicit knowledge telling than knowledge transforming from candidates, which indicates that these task types are probably best avoided in any test of writing which aims to assess advanced levels of written proficiency.

Thus far we have established that models of writing rest on models of working memory and that while there is much experimental and neuropsychological evidence for some of these, questions remain about how some features of working memory function. It has also been shown that while models of writing exist, many of them were developed with reference to native speakers. Nevertheless, if we are to explore issues of validity in writing, such models of working memory and cognition in writing will have to underpin our definition of what is involved in the production of text. The next part of this chapter now goes on to examine some of the

problems around how evidence of mental processing has been collected in studies.

2.5.2 The second problem: methodology used in studies of mental processing

A key issue for those concerned with the cognitive processes carried out by writers is the matter of how these processes are to be accessed.

Simple observation and video or keyboard tracking software can accurately demonstrate how written output takes shape but these techniques provide information on the physical process of writing and not on the mental process in which the writer is engaged. Of course observation can reveal certain features such as how and when writers correct or amend what they write but this does not reveal the cognitive process that led up to this action. Other features of composition such as consideration of the task, the organisation of the material and the on-going changes to any plan that the writer has in mind remain invisible.

The question of how to access cognitive processes has been a major difficulty for researchers in this area and two key problems have emerged. First of all, how can cognitive processes be accessed reliably and second, to what extent is there a danger that the accessing of these processes causes them to alter as unconscious processes are made conscious.

These two issues, which will now be explored, are vital in the discussion of validity and cognitive validity since those designing language assessments seek to ensure that the cognitive processes carried out by candidates mirror those (as far as is possible) that would be used in the world beyond the examination.

2.5.2.1 Verbal Protocols

Of the data collection methods into the cognitive processes of writing verbal protocols are one of the most frequently employed techniques. This method of research was pioneered at the start of the twentieth

century but fell out of fashion with the rise of behaviourism (Brown & Rodgers, 2002, p. 53) and was condemned by researchers for being “untrustworthy for scientific purposes” (ibid). Decades later Ericsson and Simon (1980) argued that many of the problems identified with verbal reports were not necessarily flaws with the method but related to the erroneous ways in which data had been collected. As a result they proposed a set of principles to guide the use of verbal protocols. Ericsson and Simon were also careful to limit the claims that could be made from such data stating that “we will not assume that the verbalised description accurately reflects the internal structure of processes or of heeded information, or that it has any privileged status as a direct observation” (Ericsson & Simon, 1980, p. 217). In other words, while verbal data might be useful for inferencing cognitive processes and model building, it does not prove that these are certainly the cognitive processes which have taken place.

This section will now define the term verbal protocol before going on to examine the two key issues related to the validity of these methods for evidencing cognitive processes.

Verbal protocols or reports (the terms will be used interchangeably here) are procedures which are used in order to infer “thought processes and attended information from behaviour” (Green, 1998, p. 4). As set out above, no claim can be made by researchers to be directly observing the actual cognitive processes that a subject is using. However, the method allows for predictions to be made about how people will carry out tasks and to what they will attend to while writing. These predictions can also be validated through the use of verbal protocols or reports. Unlike discourse analysis, verbal protocols are concerned less with the actual language of the resulting report than with what the report implies about the cognitive processes which were utilised by the subject.

Different forms of verbal protocol exist. Broadly speaking, verbal protocols are divided into tasks which are 'talk aloud' where the "information is already encoded in verbal form" (Green, 1998, p. 5) and 'think aloud' where some of the information may already be in verbal form but in which the participant will need to encode other information such as the location of an object. Both of these two tasks (talk aloud and think aloud) can be applied either to the task concurrently or retrospectively. Finally, there is a potential difference in what Green (1998, p. 5) terms 'procedural' aspects as to whether the output from the participant is 'mediated' or 'non-mediated'. A non-mediated response is where the participant reports using their own words, while a mediated response has more guidance from the researcher, for example questions such as "why did you do that?" might be used.

Objections to the validity of the use of verbal reports for the provision of evidence of cognitive processes focus on two key issues; veridicality and reactivity. Both of these issues must be addressed as these two factors potentially could undermine any evidence for cognitive processes gleaned from verbal reports. Many writers have cited these problems as strong reasons for discounting or severely limiting the value of evidence gained via such methods. Lashley (1923, p. 352) in his critique of verbal reports reduced the use of the method to the generation of hypotheses which would then have to be investigated by the use of what he considered to be more objective measures. In their examination of verbal protocols, Ericsson and Simon (1980) acknowledge the criticisms of verbal reports but point out that crucially critics have rarely distinguished between different types of verbal protocol. Ericsson and Simon argue that the criticisms of the method such as by Nisbett and Wilson (1977) focus on studies which were methodologically flawed. Nevertheless, since both veridicality and reactivity potentially pose threats to the validity of the resulting models of cognitive processing, these two issues will now be looked at in turn.

2.5.2.2 Veridicality

Barkaoui, (2011) defines veridicality as concerning “whether the think aloud protocols accurately report and represent the participants’ true and complete thinking” (p. 52). The model for cognitive processing put forward by Ericsson and Simon (1980) proposes a distinction between the operations of short-term memory (STM) and long-term memory (LTM). Ericsson and Simon argue that if a task is being reported either concurrently or immediately after it has been carried out (retrospective), then those elements to which the subject attended and heeded in carrying out the task will still be present and available for recall. Also, given the temporal proximity of the task, the LTM cannot be drawn on in providing this description due to mental overcrowding. This account is consistent with models of working memory described earlier in this chapter. However, Ericsson and Simon identify a number of essential conditions for the valid implementation of verbal reporting. Firstly, the task and the verbalisation should be interrelated so the subject must be reporting on the task that they were set, not on a different task (p.228). Secondly, veridicality becomes problematic where a subject is asked to account for their behaviour rather than simply describe it. The cognitive processes used may have been unconscious and automated so the participant may not be explicitly aware of them so in being asked to provide an explanation the participant has to draw on unheeded information. Such recollection may be not only difficult but actually almost impossible so the participant is likely to draw upon resources from LTM, reporting what they believe they did, thereby producing what Nisbett and Wilson (1977) identify as being incongruent self-explanations of participant behaviour. Ericsson and Simon (1980, p. 222) identify similar problems with the use of direct probes such as “did you use/do...” where the participant may provide the answer which they feel the experimenter wants. In short, STM will allow for description of processes but not explanation. Ericsson and Simon extend the same principles to retrospective reports (1980, p. 226) but with

the proviso that such reports must be carried out immediately after the task. Similarly, Ericsson (2002, p. 985) reports that incidents of 'forced recall' in studies by Meissner *et al* (2001, cited by Ericsson, 2002) where participants were required to speak for a specified amount of time also increased issues in reactivity as the subjects deployed LTM resources in order to fill the time they were required to do so.

So in summary, it appears that veridicality is unlikely to be an issue so long as:

- verbal reports are carried out while the heeded information is still within the working memory
- the verbal report is related to the task carried out
- the prompts used for the task require the subject to describe what they have done rather than explain.

An additional issue, related to veridicality concerns whether a participant would be able to report cognitive or metacognitive processes due to their automated or semi-automated nature. As set out in 2.4.1, metacognitive strategies may become so automated as to appear to lose the appearance of being higher level processes, despite the fact that they still control much of the process. This would be expected in the case of participants with stronger writing skills. As set out in 2.5.2.1, the use of verbal reports can only be used to infer the use of cognitive processes and this would extend to the use of metacognitive strategies.

2.5.2.3 Reactivity

The second issue often levelled at the use of verbal reports is that of reactivity or the issue of whether "the requirement to report the...process alters the process being observed or its outcomes" (Barkaoui, 2011, p. 52). A study by Plakans (2009, p. 567) of international students carrying out a writing task serves as an example of this phenomenon. Plakans

found that three of the six participants in the study reported that the process of speaking aloud while composing their essays had helped them in their thinking and assisted in their proof-reading. If this is the case then it raises issues regarding the validity of data collected in this way as the process of data collection would appear to influence both the cognitive processes and the outcome.

Ericsson and Simon (1980) argue that the process of verbal reporting, based as it is upon processes in STM, should not impact upon or change the cognitive processes that candidates use for tasks other than in some occasions extending the amount of time the production of a text takes due to the need to report back. They cite studies such as that by Karpf (1973 cited in Ericsson & Simon, 1980, p.228) which demonstrated that there was no significant difference between the performance of groups asked to carry out tasks in silence and those asked to verbally report. Ericsson and Simon do concede that the act of reporting will slow down performance; this was found to be the case in most studies but without impacting on the quality of the performance.

However, as in the case of veridicality, issues do arise if participants are asked to explain or account for their actions or if there is too great a time lapse between the task and the verbal report as the participant is then required to draw upon some aspects of LTM in order to produce their account. In his 2002 article on overshadowing (that is where the task procedures distort the verbal report), Ericsson reports that previous studies using the *Tower of Hanoi* task resulted in the task being completed in a shorter number of moves when the participants were asked to explain each move that they were carrying out (Ericsson, 2002, p. 982). As is the case in their recommendations on veridicality, Ericsson and Simon (1980) and Ericsson (2002) suggest that the probes used should be general and non-specific in order to prevent participants from having to draw on LTM or other resources.

The report of reactivity in Plakan's (2009) study above could possibly be considered as deriving from one of two sources. Firstly, the candidates were asked to think-aloud throughout the writing process, meaning that the participants were reading aloud when they were monitoring their work. This may have initiated a more careful reading than would have been done silently (Gibson, 2008). However it is **also the case** that these observations were made post-task when the participants well may have had time to reflect and reconstruct what they felt about the process **as well as having the opportunity to retrospectively consider metacognitive strategies which they had been taught to apply.** Certainly, Plakans seems surprised by the candidates' suggestion that reporting had improved their proofing as the process did not appear to be any more noticeable for these participants than for others in the study. However, it does raise the issue of whether the act of having to simultaneously write in a foreign language and report it verbally may be too much for the STM, thereby requiring candidates to draw on other resources, such as LTM.

2.6 The model of cognition in writing in the study

The previous sections have referred to a number of cognitive models of the writing process. In order to begin an investigation into the validity of written tasks it is necessary to identify which model will be adopted as this will provide the theoretical grounding for the investigation.

As has been mentioned earlier in this chapter, the model of writing put forward by Hayes and Flower (1980) was criticised for its failure to account for writing as a social act or to distinguish between the performances of more and less experienced writers. While other models (Bereiter & Scardamalia, 1987; Hayes, 1996; Grabe & Kaplan, 1996) have incorporated task demands and social context **into them**, they also face key problems. The Bereiter and Scardamalia models while accounting for the different processes of knowledge telling and knowledge transforming is

unable to suggest how a writer moves from being a novice to an 'expert' writer. An issue it shares with the Grabe and Kaplan and Hayes and Flower (including Hayes' later amendments) is that the model does not account for the actual cognitive process of writing. Essentially the model proposes what processes are engaged but not how these interact in order to generate a composition. The Grabe and Kaplan model also does not sufficiently separate elements of Long Term Memory from those of Working Memory.

The model provided by Field (2004) which is based on the work of Kellogg (1999) demonstrates a number of features to recommend it as a model that could be applied for the purposes of this study. Firstly, the model is rigorously underpinned by psycholinguistic theory (Shaw & Weir, 2007, p. 37) as it draws on the model of working memory put forward by Gathercole & Baddeley (1993). Secondly, the model proposes *how* composition is carried out and offers a prediction of what stages writers will go through and how these processes might vary between weaker and stronger writers; thereby suggesting a model of how long-term memory is drawn upon by working memory processes.

2.7 The model proposed by Field

The cognitive model put forward by Field (2004) and adapted in Shaw and Weir (2007: p.37) is set out in Table 2.1 below. As with other models and in keeping with Hayes and Flower (1980), Field identifies that the process is not a linear one but recursive. The model is based on Kellogg's (1999) proposed cognitive components of writing skill (see Figure 2.1) and accounts not only for the stages of writing (or 'phases' as Kellogg terms them (1999, p. 26) but also illustrates how working memory interacts with the long-term memory during the writing process.

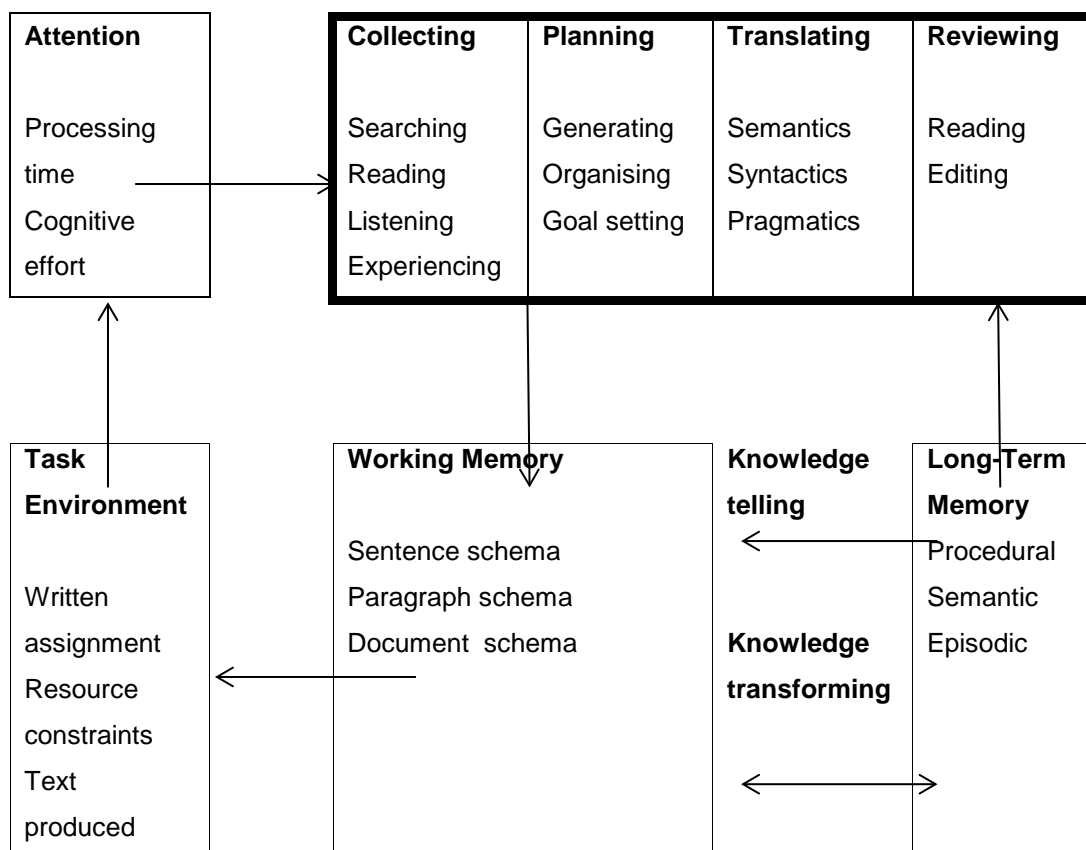
Table 2.1. *Cognitive processing framework (Shaw and Weir, 2007, p.37-43, based on Field (2004, p.329).*

Stage	Definition
Macroplanning	Pulling together ideas/resources for task. Identifying features (genre, intended audience, purpose)
Organisation	Sequencing/prioritising ideas based on decisions made above
Microplanning	Focus on section of text immediately to be produced with consideration to all previous decisions made
Translation	Turning content from mental (private) thoughts to public text
Monitoring	Examining text for accuracy and adherence to intended purpose/audience etc)
Revising	Returning to text to amend following monitoring

Field (2004, p. 329) includes an additional ‘execution’ stage in his model of the process which is omitted from the Shaw and Weir version. This stage follows on from translation and is the actual physical act of writing and presumably it has been removed from the Shaw and Weir version as it appears to sit outside the cognitive model. However, Field (2004, p. 330) makes the point that for many writers actual execution is the start of the process, which then engages the other cognitive phases. This seems to be the case also with people writing in English as a second language where writers often spend less time planning or considering the goals of their piece of work (Hyland, 2002, p. 26). Kellogg (1999, p. 28) also identifies that a number of writers such as Nystrand (1982) have argued that the distinction between the planning, translation and by extension execution stages are unnecessary. The objection raised is that the separation of these stages originates from studies on spoken production whereas written language can allow the linguistic resources to shape the planning. Kellogg acknowledges this issue but still sees validity in

separating out the stages as he argues that inner speech often precedes the translation and execution phases.

Figure 2.1. *Cognitive components of writing skill* (Kellogg, 1999, p. 26)



The lack of planning identified by Hyland (2002) when considered alongside the model of the **cognitive** process proposed by Kellogg (1999) also assists in providing a rationale for the practice of knowledge telling by language learners at the the lower levels of linguistic competence. Such learners are more concerned with the content (the *what* of their text) and the actual words and grammar they will use (*how they will say it*) than with notions of audience or the purpose of the text. To put it in other words, there simply is not enough processing capacity in the working memory to manage the arrangement of text alongside the focus on rhetorical problems. Therefore the learners will draw ideas from the long-term memory until they run out of them without considering where these might fit best in the text.

We will now turn to Fields's (2004) model which is proposed as the model which will be used to examine the validity of the writing tasks in this study. At this point each part of the model will be examined. However, certain caveats must be made regarding the model. First of all, Kellogg (1999, p. 26) states that these are *phases* not a process. This distinction is important because describing the model as a process would imply that its implementation conformed to a particular order and as has already been pointed out, the actual ordering of the phases may well be non-linear and some phases may be prolonged, recursive or even omitted entirely, particularly in the case of less experienced writers.

A second caveat is that the actual boundaries between the different phases of the model are, in the words of Kellogg, "fuzzy" (1999, p. 27) due in part to the way in which phases may overlap or be triggered by each other. For example, a writer who is in the act of translating, and actually executing the writing, may find that a particular idea occurs which results in a new organising phase. In such a situation it would be very hard to say exactly when translating became organising.

2.7.1 Macroplanning

The first phase in the model proposed by Field is that of macroplanning. As explained above, it is not assumed that writers start their compositions with macroplanning. Field (2004, p.330) points out that some writers start by immediately producing text rather than by carrying out the activities associated with macroplanning.

In the macroplanning phase the writer pulls together the ideas required for the task and the goals of the piece of writing from their personal knowledge and experience including their knowledge of the genre and expectations of the reader. Field (2004, p. 329) includes considerations of style such as the level of formality required. The knowledge of genre that the writer draws upon may be based on experience gained from similar

previous tasks and feedback or from reading of the type of text to be produced. Such knowledge will be drawn from long-term memory and existing schema. Depending on the extent of a writer's experience, he or she may need to use what Gathercole and Baddereley (1993, p. 5) term the supervisory attentional system (SAS). For example, when producing a genre for the first time, the writer may encounter a number of problems reconciling previous schema with the type of text to be produced. The model suggests that experienced writers who are knowledge transforming will also probably employ the SAS in order to problematize the writing task to ensure that the response is as effective as possible. Less experienced or less-proficient writers who are more likely to knowledge tell may well simply employ a simple pre-existing schema in order to carry out the task with only a minimal consideration of genre and task. These writers will probably use the resources of their working memory to focus more on the actual microplanning of what they are writing and the translation of internal thoughts into text.

In terms of timed essay writing, the type of writing considered in this study, the macroplanning phase may be largely unobservable in that the activity may be entirely internal with the inner voice putting together resources in a highly abstract form (Kellogg, 1999, p. 29). There may be some physical activity, with key words or terms being jotted down in a pre-planning form or some form of notation. Kellogg (ibid) points out that at this stage such notation, if it occurs, may be in the form of non-consensual symbols in that they are not intended to be understood by anyone other than the actual writer and do not form part of the communication.

2.7.2 Organisation

The organisation phase, similar to the macroplanning phrase, may also be highly abstract and unobservable. The phase sees the content generated arranged in relation to the genre and task itself but the organisation may also be based on the relationships of the ideas generated previously

(Field, 2004, p. 329). The use of non-consensual symbols is often a feature of this phase as any annotations made are not usually intended to be perceived as being part of the resulting text.

The cognitive processing which takes place while a writer is in the organising phase will depend upon the degree of familiarity of the task. The model offered by Kellogg (1999) suggests that writers who are producing genres and tasks with which they are familiar will be able to make more use of pre-existing schema leaving the working memory more space to consider the audience and specific requirements of the task.

The level of language ability is another factor which impacts on the ability of the writer to carry out effective organisation of the text. Shaw and Weir (2007, p. 38) propose that a writer who is struggling to produce the language for a task will have fewer cognitive resources to devote to organisation. This implies that with the working memory focussed on language production, organisation would either be haphazard, i.e. as ideas occur, or would rely on a pre-existing schema which may not be suitable for the given task. An example of this would be a candidate writing an advantages/disadvantages response to a question requiring a compare/contrast type answer. This overloading of the working memory would provide another explanation for the differences between those engaged in knowledge telling and those who are knowledge transforming.

With regard to timed essay writing, and indeed to all other types of writing, effective organisation is dependent upon the nature of the set task. Candidates are expected to be able to analyse the topic of the essay, identify the required response in terms of rhetorical organisation (e.g. cause-effect, compare-contrast, advantage-disadvantage etc.) and structure their response according to the demands of the genre, although as Kellogg points out this is not a sequential process (1999). The question set by the task is therefore extremely important because it will provide the

framework for structuring the response. The macroplanning phase of writing is the phase in which a candidate marshals their knowledge and arguments whereas the organisation stage will see this information placed into the appropriate shape in terms of paragraphing and sequencing of content.

It is very common to find teachers and authors extolling the benefits of learners engaging in formalised pre-writing planning; i.e. composing a plan on paper. Ellis and Yuan's study (2004) suggested that for 42 Chinese students tasked with composing a narrative, planning appeared to produce a significant improvement in fluency and syntactic complexity. However, other writers have suggested that the gains made from pre-task planning may be quite small (Johnson, 2012) or that the process may even be detrimental to the learners' performance (Ong & Zhang, 2010). This latter study is of particular relevance because it focussed on the production of argumentative tasks, similar to the task type under consideration in this study.

Johnson (2012) proposes that one reason for the different outcomes in the studies looking at the impact of planning on writing could be that there is a minimum threshold not simply of linguistic proficiency but of writing ability at which planning becomes an effective process. Johnson (2014) identifies that factors such as genre knowledge and explicit instruction in the features of the genre are likely to have an impact on the success of the planning process.

The issue of the extent to which pre-task planning assists language learners is also one which has been considered by examination developers. One of the features of the Cambridge ESOL examinations which Shaw and Weir (2007, p. 45-53) explore is the degree to which candidates are required at different levels of the CEFR to structure their own text. They observe that even up to the B2 level the examinations provide guidance in the structuring of the text within the task instructions

and that at the A1 and A2 levels this aspect of writing is more or less entirely provided for the learners. Therefore it can be concluded that from the B2 level and above, responsibility for the macro-structuring of texts is largely left to the candidate to develop.

2.7.3 Microplanning

Microplanning is the phase in which the writer's attention moves to the section of the text which he or she is about to begin producing (Shaw & Weir, 2007, p. 39). The phase is both forward and backward looking, considering what is to come in the light of what has been planned and produced already. The writer must consider both the sentence to be produced and its role at the local level of the text (i.e. the paragraph) as well as how the developing paragraph contributes to the global shape and purpose of the text.

Field (2004, p. 329) emphasises the importance of the microplanning phase for the production of the actual language to be used. Issues such as whether information is given or new (i.e. has it been alluded to in the text previously or is it information which can be assumed to be shared with the reader) impact on the structure of sentences and the way in which ideas are communicated. Competent writers will bear in mind the macro-level notions of the text purpose and the intended audience while making decisions about how to convey the information. It is therefore logical to assume that candidates who are linguistically weaker, or less competent at writing will once again face an overload in their working memory as their mental resources are more likely to be taken up by producing language rather than considering the reader's assumed knowledge. Shaw and Weir (2007, p. 39) identify the failure to take into account the role of text produced so far as a problem in the cognitive model of writing put forward by Grabe and Kaplan (1996). This is an important omission because Field's model suggests that these phases are reoccurring and that competent writers will continually be reviewing the text which they have

produced in order to assess whether it is keeping to the intended purpose as well as demonstrating local coherence and cohesion. It can be inferred that, if Field's model is correct, those candidates who are linguistically weaker or less-capable as writers will tend to focus on the micro-level of the immediate sentence without considering the links to the more global phases such as macroplanning and organisation.

2.7.4 Translation (and Execution)

The translation phase is where text is converted from abstract mental ideas to physical written text. In other words when thoughts or non-consensual notes are turned into consensual script. Field (2004, p. 329) describes the process of making use of what he terms a *writing buffer*, whereby the writer knows when they start a sentence how they will finish it. The concept of the phonological loop (Gathercole & Baddeley, 1993) is drawn on by Kellogg (1999, p. 29) and is useful in the discussion of the *writing buffer* and the use of working memory, as it is the use of the 'inner voice' to code thoughts into writing. Evidence of the phonological storage of the information is often presented through examples of errors of substitution, where words like *there* and *their* are mistakenly replaced or even whole expressions: *as a pose to* instead of *as opposed to*.

For those writing in a language other than their own, translation is the point where there is the potential conflict between what a candidate wishes to convey and the linguistic resources that they have available to them. Field (2004) identifies that those writing in a language other than their first language will often employ strategies such as avoiding certain structures which they are not confident of or using language that they do have which is not strictly accurate to convey their ideas by means of circumlocution to express their meaning.

Shaw and Weir (2007, p. 39-41) omit the execution stage which Field (2004, p. 329-330) details, preferring to focus on the cognitive process of

writing. Execution may well have a considerable role in cognitive processing, especially for competent writers who may often appear not to macroplan or organise their text but simply begin writing. Field suggests that the physical act of writing may, for these individuals, trigger the planning phases.

2.7.5 Monitoring

Monitoring as a phase is constantly employed by skilled writers since throughout the phases of writing (macroplanning, organising, microplanning, translating, editing and revising) there is the need to consider the purpose and audience of the text as well as the effectiveness of the part of the text under construction in contributing to that purpose. Field (2004, p. 330) proposes that only one level of monitoring can be carried out at a time whether this be attention to the sentence under construction or overall concern for the text as a whole. This one-level at a time approach is unsurprising given studies like those of Roussey and Piolar (2008, cited in Eysenck & Keane, 2010, p. 447) who identified that reviewing placed additional strain on the working memory and was more challenging for those whose working memory had a low capacity.

Often, during the writing process, monitoring will be on the micro-level of the word, clause or sentence being produced. Eysenck and Keane (2010) suggest that those writing in a foreign language will tend to fixate on linguistic features rather than issues of organisation, discourse, audience or genre. Again, this is unsurprising given what has been said about the demands placed on working memory by the monitoring process and the cognitive load imposed by the linguistic demands of a task.

2.7.6 Revising

Alongside monitoring, revising may take place during any phase of writing though studies by Chenoweth and Hayes (2001, p. 94) suggest that second language users are more effective at editing their tasks post-

composition rather than during. The authors hypothesise that this is due to the interference from meaning-focussed attention while they are writing. Various writers (Hayes & Flower, 1986; Field, 2004; Shaw & Weir, 2007; Eysenck & Keane, 2010) identify that more proficient writers (including L1 writers) tend to spend longer editing their texts than less-skilled authors and that those writers with higher skills will tend to make changes which are meaning-focussed rather than simply surface level changes. Field (2004, p.330) suggests that these changes often occur at the lexical level but that the changes operate on different domains of the text; for example one word may be substituted for another because the tone and appropriateness of the alternative are felt to be more in keeping with the genre and intended audience of the text; for example changing *things* for *factors* in an essay.

Those writing in a second language are often found to be preoccupied with surface level linguistic changes to a text but this could also be attributed to the types of feedback which they receive from teachers which often focuses exclusively on this level (Cohen & Cavalcanti, 1990).

2.8 Studies of cognitive processing in L2

Studies which have explored cognitive processing in L2 test-writing through the use of verbal reports have identified a number of key elements. Mickan, Slater & Gibson (2000) and Yu, Rea-Dickins and Kiely (2007), investigating the IELTS writing examinations and Chan (2011), writing about the Pearson Academic English Test, all identified the impact of the task as being an essential element in the eliciting of cognitive processes by test takers. Mickan, Slater & Gibson point out that a candidate's interpretation of the task will impact on planning and shape the response to the task a finding which Yu, Rea-Dickins and Kiely concurred with, emphasising the importance of task familiarity in shaping a writer's response. Chan, contrasting the cognitive processes elicited by writing only tasks with reading into writing tasks noted that essay writing tasks on

their own tended to encourage knowledge-telling responses. Participants in Chan's study who carried out the writing only task spent minimal amounts of time macroplanning, and tended to generate content from searching their own experiences and knowledge and immediately microplanning and producing text in a hand-to-mouth fashion. As a result, there was little text level organisation. Chan identified that these participants did spend a fair amount of time editing their work but that it remained at a rather superficial level, similar to Cohen & Cavalcanti's observation (1990).

Micken, Slater & Gibson (2000), who used nine B2 level students (IELTS 6) also reported little evidence of advanced planning among these learners (p.46), similar to Plakans (2008), which may also add some support to Johnson's argument (2012) that there may be a minimum level at which planning becomes effective. This view concurs with that of Manchon and Roca de Larios (2007) who found that level of language proficiency correlated with the amount of time devoted to planning a piece of timed writing. Roca de Larios et al (2008) also identified that the level of proficiency in language of a writer affected the distribution of cognitive processes across the initial, middle and final period of writing time.

Another element that Micken, Slater & Gibson (2000) and Chan (2011) identify is that participants in their studies did make use of metacognitive knowledge of the essay genre in order to structure their texts at the broadest level (introduction, main body, conclusion) however there seems to have been little consideration of how genre might impact on microlevel decisions during writing or on appropriate language selection with text production described as being 'linear' in nature.

The studies outlined above suggest that any investigation of cognitive processes used in timed writing must take into account the level of proficiency of those taking part as this will impact on the processing

carried out. The task set is also of crucial importance and that timed essay writing on its own may favour knowledge telling linear responses by learners. The studies also predict that candidates at lower level will make less use of planning time and macro-level planning metacognitive strategies.

2.9 Conclusion

This chapter has set out to explore the issue of how arguments for validity are key to the development of a test. We have seen that validity is a multi-faceted concept and that any claim must be supported by evidence and is not a one-off step. In the case of writing, a crucial area for examination is the question of whether a claim of cognitive validity can be made; that is that the cognitive processes which test-takers engage in while completing a task elicit the same cognitive processes that tasks in the real-world will require.

Chapter Two has established that there are two methods for exploring the issue of validity: *a priori* and *a posterior* (Weir, 2005). While *a priori* investigation are often associated with test design and *a posterior* with post-test validation, both can make use of product and process orientated-methods. A *product*-orientated method is one which interrogates the texts produced while a *process*-orientated method attempts to infer the cognitive processes which writers engage during the act of writing. This chapter has focussed on this process-orientated dimension and argued for the use of verbal protocols to explore how candidates compose their texts. Models of cognitive processing in writing which draw on psycholinguistic studies have been considered to provide a framework against which the processes elicited from candidates in the test can be compared. Chapter Three will consider how validity is argued in writing tasks through the use of criteria based on theoretical models. These models are used by criteria to represent what effective writing is considered to be in terms of test performance and embody the construct of writing against which a

candidate is being measured. The chapter will then consider the role of discourse competence as a key component of such models of writing and as an indicator of proficiency at higher levels. The discussion of discourse competence will seek to establish what features in texts produced by candidates could be used as evidence of proficiency at the levels of B2 and C1.

Chapter Three: Writing and Communicative Competence

3.1 Introduction

This study aims to investigate two key areas in relation to the ESB ESOL International Examinations, namely:

1. To what extent is **cognitive validity demonstrated** in the cognitive phases that candidates carry out while producing scripts in the English Speaking Board ESOL International Examinations?
2. What is the role of discourse competence in deciding whether a script is classified as being level B2 and C1 of the Common European Framework of Reference for Languages (CEFR) in candidate scripts from the ESB ESOL International Examinations?

The previous chapter has identified that validity is a multi-faceted concept and has explored the concept of cognitive validity in relation to the production of writing by language learners. This next chapter now explores how validity is related to the underlying constructs and theories that underpin assessments. The argument will run that in order for a test of writing to build a case for validation, it must be constructed around theories of what constitutes good writing and of how candidates at different levels produce text. Chapter Three sets out how discourse competence can be used to explore candidate performance at the CEFR B2 and C1 levels of proficiency by building a link between the models of communicative competence set out in the literature and the CEFR.

3.2 Validity in *a priori* assessment design

Building an argument for validity in the development of assessments must begin with the theoretical underpinning of the method of assessment. The designers of assessments must show how the test links with the

knowledge, understanding or performances under examination. The link between the underlying constructs and the assessment is essential to establish if there is to be confidence in the assessment, and for an argument for validation to be made. Weir (2005, p. 18) makes the case for clear theoretical definitions of what is to be assessed right from the start of the development of an assessment. It is therefore necessary at this point to explore the theory which underpins communicative tests of English writing such as those being investigated in this study.

3.3 The concept of communicative competence

Communicative competence is defined by the CEFR (Council of Europe, 2001, p. 9) as being the competences “which empower a person to act using specifically linguistic means”. The CEFR identified these as being linguistic, sociolinguistic and pragmatic competences (p.13). The communicative paradigm within which much of English Language Teaching takes place has been in existence since the 1970s and is founded on the concept of communicative competence. Therefore, in this context of language teaching, an assessment must be linked to a communicative model of language and language use in order for a claim of validity to be made. Such a model must account not only for linguistic features (i.e. the actual language used) but must also incorporate sociolinguistic and pragmatic features. The next section considers the origins of different models of communicative competence and the theory which underpins them.

3.4 The origins of communicative competence

The origins of communicative competence lie in Chomsky's (1965) distinction between competence, which is an individual's knowledge of the language, and performance; the actual use of language by an individual (Richards, Platt, & Weber, 1985, p. 211). Competence was viewed by Chomsky as an unchanging state (Celce-Murcia, Dornyei, & Thurrell, 1995, p. 7) and was distinct from proficiency. In Chomsky's model,

competence was the abstract knowledge of the language including grammar and assumed no constraints such as time or memory pressures. Essentially Chomsky's model of competence is an individual's knowledge of the grammatical systems of a language under ideal conditions. Linguistic performance was viewed as a process and is limited by factors such as time, cognitive demands, social situation and other contextual elements.

Chomsky's (1965) distinction between competence and performance was challenged by Hymes (1972) who argued that the role of context was ignored in Chomsky's distinction and that socio-linguistics played a vital role in determining what could be said and how it could be conveyed in different contexts. Hymes also claimed that the role of social context was central to language and in particular the knowledge of how to use language in different social situations. Hymes stated that "there are rules of use without which the rules of grammar would be useless" (p.278). The concept of communicative competence therefore encompasses not only grammatical syntactical competences but a range of sociolinguistic and pragmatic competences.

Another important influence on the emerging concept of communicative competence was the work of Halliday and Hassan and their work *Cohesion in English* (1976). Halliday had already put forward the view that text was a product of context, or the context of situation as proposed by Firth (1950) who had also influenced the work of Hymes (1972). McIntosh, Halliday and Stevens (1966) devised a system for the identification of register through the examination of the Field, Tenor and Mode of a text. These three notions overlap to some extent with the three functional-semantic components: the ideational, interpersonal and textual. The ideational represents the content of a text; the business that it is carrying out and the message which it is conveying. The interpersonal is connected to the relationships of those involved in the communication, the

‘angle’ (Halliday and Hassan, 1976, p.27) of the writer/speaker and their motives while the textual function is to do with the linguistic resources used for the purpose. Halliday’s model of language and the emphasis on context proved to be another important influence on the development of the notion of communicative competence.

The development of communicative competence as a concept had a major impact on language teaching because it emphasised the vital shaping role of context on language and the indivisibility of factors such as text purpose, intended audience and social-factors. Clearly, if communicative competence is the model upon which language learning and assessment is based then the validity of an assessment will be determined by the extent to which it represents the model in its constructs.

3.5 Models of communicative competence

This section now considers the major models of communicative competence based on the concepts set out in the previous section and considers how these models incorporate the different competencies: linguistic, socio-linguistic and pragmatic. Understanding of these models and of their components is essential in order to be able to provide a coherent link between the theory which underpins an assessment and the building of *a priori* and *a posteriori* arguments for validation.

3.5.1 Canale and Swain (1980)

The model of communicative competence developed in 1980 by Canale and Swain has been particularly influential in that it is often regarded as being the cornerstone of communicative approaches. It has also influenced the CEFR (Council of Europe, 2001) and is referenced in the document. Fulcher (2010: p.106) points out that Canale and Swain’s model was intended to assist those attempting to develop syllabuses and examinations by developing a model of communicative competence to underpin the emerging communicative approach.

Canale and Swain's model (1980) views communicative competence as being made up of three competences: grammatical competence, sociolinguistic competence and strategic competence. Grammatical competence overlaps with Chomsky's definition of competence in that it comprises "morphology, syntax, sentence-grammar semantics and phonology" (Canale & Swain, 1980, p. 29). Socio-linguistic competence was made up of two elements: sociocultural rules and rules of discourse (Canale & Swain, 1980, p. 30). Socio-linguistic competence therefore is concerned with the context appropriate understanding of utterances dependent upon the roles and relationships of those involved, the settings and the norms of the situation.

Strategic competence was seen by Canale and Swain as being a new element that they had brought to the model of communicative competence. They stated that aside from one or two exceptions in the literature, there had been little consideration of how speakers deal with breakdowns in communication (e.g. false starts, hesitations and performance errors) or manage uncertainties (e.g. being unsure how to address someone in a particular context). So strategic competences comprises the strategies that speakers use to maintain the flow of communication (Canale & Swain, 1980, p. 25).

While Canale and Swain's (1980) model of communicative competence is divided into different elements, it was also to be understood that these were connected and influenced each other; for example socio-linguistic awareness would influence grammatical choice, while grammatical competence would influence the range of cohesive devices that a writer or speaker might be able to deploy.

Discourse competence was acknowledged by Canale and Swain (1980, p.20) as an important part of the system of communicative competence,

but they also stated that there was “no theory of discourse that one can turn to with confidence”. They based their view of discourse on the work of Halliday and Hassan (1976) and therefore considered discourse to be part of sociolinguistic competence (1980: p.30) made up of coherence (“appropriate combinations of communicative functions”p.30) and cohesion (grammatical and lexical links). They acknowledged that this was a pragmatic decision “until more clear-cut theoretical statements about rules of discourse emerge” (ibid). Clearly though, discourse was an important component in the creation of a text. Chomsky (1965) had established that written language could be grammatically accurate and yet lack the qualities to be considered a text, what Halliday and Hassan (1976, p. 2, 23) describe as being “a unit not of form but of meaning...realised by, or encoded by sentences....it [a text] is coherent with respect to the context of situation and therefore consistent in register; and it is coherent with respect to itself, and therefore cohesive.” It can be seen from this that coherence as a concept was identified by Halliday and Hassan as having strong connections with the interpersonal functions of the text in that it had to be consistent with the required register while cohesion is a feature of the textual functions of a text but should be regarded as being non-structural (p.29) in that cohesion is a feature that operates both within and beyond the sentence or clause level of a text.

3.5.2 *Canale (1983)*

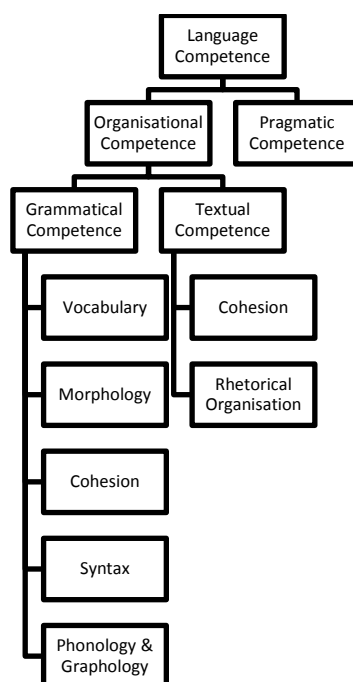
In 1983, Canale significantly altered the model of communicative competence proposed by Canale and Swain (1980) by separating out the area of discourse competence from socio-linguistic competence. Canale maintained the distinction Canale and Swain had made between communicative competence and performance but relabelled the psychological and environmental conditions which might impact on production as ‘actual competence’ (Canale, 1983, p. 5).

Discourse competence was identified by Canale (1983, p. 9) as being the force that allows a speaker (or writer) to produce a particular genre of text (again, whether written or spoken) in a unified way. The key principles of coherence and cohesion were again important in this model (1983, p. 9-10) as the methods by which textual unity was maintained. Canale also demonstrated interactions between grammatical, sociolinguistic and discourse rules and the influence of these areas upon each other, again emphasising the impact of different features upon each other.

3.5.3 *Bachman (1990) and Bachman and Palmer (1993)*

The model of communicative competence proposed by Bachman (1990) and further developed by Bachman and Palmer (1996) was devised as a model which would assist those designing language assessments. One of the innovative features of this model was the division of knowledge and skills, which Canale (1983) had not distinguished from each other. In Bachman and Palmer's model cohesion was a feature of both grammatical competence but also of textual competence. This latter competence was the equivalent of what Canale and Swain (1980) had termed discourse competence but it consisted not only of cohesion and the creation of unity within texts but also of textual and rhetorical organisation. Bachman and Palmer's system makes clear the link between discourse competence and its realisation through language while emphasising its considerable role in the overall design of a text. The Bachman and Palmer model also went further than Canale had done in exploring the cognitive elements of communicative competence. These were termed metacognitive strategies and were the processes of assessing the requirements of a particular text, goal-setting, planning and execution. As discussed in 2.4.1, the model was acknowledged to go beyond previous models in terms of applying metacognitive strategies to all areas of language use as well as exploring how the strategies interact with lexico-grammatical knowledge, Bachman and Palmer's model has been criticised for its lack of an empirical basis (Purpura, 1999).

Fig. 3.1. *Bachman's Components of Organisational Competence (Based on Fulcher, 2010, p.109).*



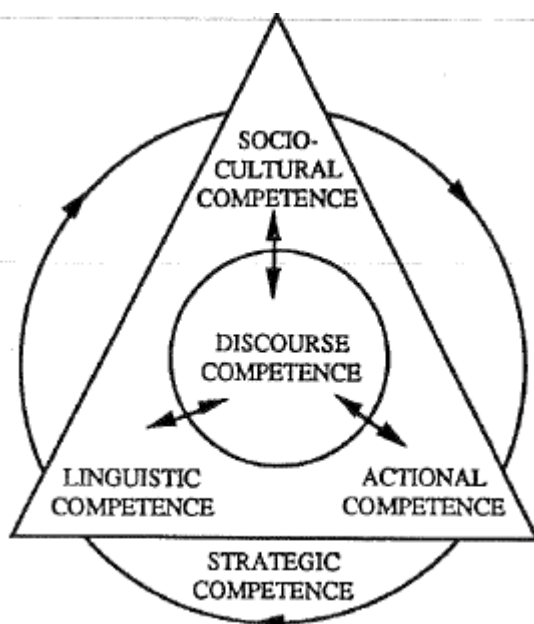
While the Bachman (1990) and Bachman and Palmer (1993) models developed the concepts proposed by Canale and Swain (1980), writers such as McNamara (1995) have identified that the model is superior for its division of knowledge and skills. However, McNamara also identified issues with the model, such as the overlap between knowledge and skill (for example the illocutionary component and the strategic components where the first is presumed to be knowledge and the latter a skill). Also the decision to separate lexis from grammatical competence by placing lexis into pragmatic competence assumes a clear division between the two which many writers such as Halliday (1985), and Celce Mercia, Dornyei and Thurrell (1995) would contest.

3.5.4 *Celce-Murcia, Dornyei and Thurrell (1995)*

The model of communicative competence proposed by Celce-Murcia, Dornyei and Thurrell continued to build on the work of Canale and Swain (1980), Canale (1983) and Bachman and Palmer (1990, 1993). The

model was designed to further develop the notion of sociolinguistic competence in that it continued to separate discourse competence from sociolinguistic but also included actional competence which is “competence in conveying and understanding communicative intent by performing and interpreting speech acts and speech act sets”. In other words actional competence concerns the language user’s ability to carry out and interpret language functions. The authors make the point that learners may be able to carry out functions even if they are not aware of the correct sociolinguistic conventions.

Figure 3.2. *Schematic representation of communicative competence* (Celce-Murcia, Dornyei and Thurrell, 1995, p.10)



While each of the models discussed above vary the way in which they divide up communicative competence, all of the models emphasise the interaction of the different elements in producing effective and appropriate communication. To return to the issue of writing, it is clear that any model of writing and any assessment of writing derived from such a model must take into account the different aspects of communicative competence.

A second point that needs to be made based on the discussion of the different models of communicative competence is that all of the models

have included discourse competence as a key component. While initially identified with the notions of coherence and cohesion in Canale and Swain's model, the later models extended the idea to include knowledge of genre. The models have also identified the link to language use, in that it is through the organisation of the text on macro and micro levels that discourse competence is achieved. It is to the role of discourse competence in the different models that we now turn in order to consider the importance of this feature in determining the development of language proficiency in the CEFR.

3.6 Discourse competence

The previous sections of this chapter have explored models of communicative competence. This exploration has been carried out because these models underpin communicative language testing and form the constructs upon which language tests and tests of writing must be based. We have established that the concept of discourse competence is shared by many of these models and that this competence provides a link between the macro-level of text design (i.e. organisation of the whole text) and the micro-level (linguistic features of cohesion). This section now looks at defining discourse competence and the role that discourse competence has been given in the definition of levels of proficiency by the CEFR. We will also discuss how discourse competence can link to Weir's model of *a priori* and *a posteriori* validity. The next section considers the role of discourse competence within the models of communicative competence already discussed before examining how the CEFR incorporated discourse competence into its model.

3.6.1 Definitions of Discourse Competence

English language teaching has largely adopted the description of discourse from work carried out in linguistics, in particular the work of Hymes (1972) and Halliday and Hassan (1976). It is important to identify

at this point what is meant by 'discourse' and how writers in the original area used the term 'discourse competence'.

Writers in the field of linguistics acknowledge that discourse is a vast area of work (Schiffrin, 1994, p. 5) within which there are different traditions and approaches. The work of Hymes (1972) originates from the area identified as the ethnography of communication and has its origins in anthropology (Schiffrin, 1994, p. 8). Hymes was unhappy with Chomsky's (1965) sidelining of 'performance' and sought to refocus study on communicative competence and how social, psychological and other factors influence communication. Hymes also made the point that what is classified as communication could vary between cultural contexts. Hymes also falls under the 'functional' approach to discourse (as indeed do Halliday and Hassan (1976)) as opposed to the 'structural' approach (as Chomsky might be characterised). Schiffrin (1994, p. 22) sets out two features which distinguish the functional approach; firstly that the functional approach takes the view that "language has functions that are external to the linguistic system itself" and secondly that the linguistic system will be in part be influenced in response to external functions. In terms of examining discourse competence, this distinction is important as a formalist, such as Chomsky, would argue that external factors would have little or no bearing on how a text was constructed in terms of coherence and cohesion, while the functional approach proposes that these external influences are central. To return briefly to the aims of this study, if one of the hallmarks of a successful candidate at the higher levels as identified by the CEFR is the ability to adapt their writing towards the expectations of the assumed reader then it follows that the CEFR adopts a functional approach to discourse and discourse competence.

The approach taken by those using a functional approach to discourse suggests a view of language which is interactional (Schiffrin, 1994, p. 401). In this view of language context is a key element to be considered

and it is the use of contextual clues to inference intention and the reading of discourse which is central in this approach. This has connections with the CEFR and its assumptions about the nature of language. The CEFR's view of discourse competence will be discussed in section 3.6.2.

According to an interactional view of discourse, more proficient users of language should be better able to take into account shared experience and knowledge in order to tailor their texts. Therefore these proficient users would be expected to leave more clues within their texts about the purpose and nature of the communication as well as the state of the interpersonal elements and the degree to which textual elements for coherence and cohesion need to be employed.

That the CEFR takes a functional approach to language is not surprising. One of its earliest forms was Van Ek's work on the *Threshold Level* (1979). The document was an important step in the development of communicative competence but also in the establishment of discourse competence as a key feature in its own right. In the Threshold document, Van Ek identified six competences which comprised communicative competence including discourse competence. Van Ek defined discourse competence as "the ability to use appropriate strategies in the construction and interpretation of texts, particularly those formed by stringing sentences together" (van Ek, 1979: p.8). In the same document he further defined discourse competences as consisting of features such as 'moves' (ibid, p.9) as well as lexical and grammatical cohesive devices, similar to those proposed by Halliday and Hassan (1976). The inclusion of 'moves', the macro-level organisation of the content to assist the communicative purpose of the text, evidences that Van Ek also saw discourse competence as operating at different levels within texts simultaneously.

Van Ek's (1979) work influenced the later models of communicative competence, including that of Canale and Swain (1980). As is set out in 3.5.1, the notion of discourse competence in the work of Canale and

Swain (1980, p.30) was originally subsumed within socio-linguistic competence. In their initial model of communicative competence, discourse competence was identified as being the rules of coherence and cohesion. The authors pointed out that there were overlaps with grammatical competence but that discourse competence was more concerned with the “combination of utterances and communicative functions” as distinct from whether a piece of text adhered to the principles of correct grammar or whether it was sociolinguistically appropriate for a given context. As has been stated earlier, the position taken by Canale and Swain was to some extent a pragmatic position due in part to what they perceived as being a lack of clarity on the rules of discourse.

In his 1983 article, Canale modified his and Swain’s (1980) earlier model of communicative competence and identified discourse competence as a separate component, distinct from socio-linguistic competence. The revised model maintained that discourse competence comprised grammatical and lexical features of cohesion and coherence, the latter linked to unity of form and the latter to unity of meaning (Canale, 1983, p.9). Canale also identified that such features would be related to the particular genre, whether it was written or spoken. In his article, Canale argued that discourse competence could be considered to be a separate area because it could be clearly differentiated from grammatical competence and socio-linguistic competence when violations of it are identified. Canale cites the following example from Widdowson (ibid, p.10):

A: What did the rains do?

B: The crops were destroyed by the rain.

Canale identifies that although from a grammatically and socio-linguistic point of view B’s response is not erroneous, it does not fit with the pattern of the discourse due to the ordering of the new and shared information. In the 1980 article, Canale (1983, p. 10) asserts that aspects of discourse

competence interact with grammatical and socio-linguistic features and do not particularly need to be separated. The separation of discourse competence from socio-linguistic competence in particular has been a point of criticism for Canale's model. Schachter (1990) argued that the construct of discourse competence was ill-defined and that if it is concerned with the larger structure of texts, then it belonged in socio-linguistic competence since discourse competence was concerned with unity of text which involves appropriateness, the status of the participants and whether the required norms and interactions are successfully carried out. However, it seems clear that Canale's model acknowledges that there are points of inter-relation between the components of Communicative Competence and that any discussion of discourse competence will necessarily have overlaps with other competences.

Despite some difficulties, Canale's (1983) definition of discourse competence is particularly useful in this study given what will later be said about the links between the forms and meanings that he identified. It is also a useful starting place as Canale's conceptualisation of discourse competence also supposes the possibility that different genres and text types will necessitate different cohesive and coherence features.

The Celce-Murcia, Dornyei and Thurrell model (1995) is presented by its authors as an evolution of the Canale and Swain (1980) and Canale (1983) models as well as the Bachman (1990) and Bachman and Palmer (1993) models. In the Celce-Murcia *et al* model, discourse competence is placed at the centre (see figure 3.2) as the meeting place of lexico-grammatical competence, actional competence and sociocultural competence, depicting the way in which discourse shapes each of these and is in turn shaped by them.

Celce-Murcia, Dornyei and Thurrell (1995) define discourse competence as "where the bottom up lexico-grammatical microlevel intersects with the

top-down signals of the macrolevel of communicative intent and sociocultural context to express attitudes and messages, and to create texts.” (1995, p.13). This definition is consistent with previous models which linked the macro and micro levels of discourse competence. The definition is also of interest because it implies that discourse competence will be demonstrated in two ways. First of all, discourse competence will show the ability of the writer to create a text that can fulfil the intended purpose by utilising the conventions of the genre and the discourse community; the “top down signals” mentioned above. Secondly, the actual use of language in the text will demonstrate the writer’s ability to use cohesive and coherence devices in their text in order to fulfil the intended purpose of the piece of writing. Since this definition encapsulates much of what this study is attempting to demonstrate, it is the definition which will be taken forward to the following discussion on the CEFR and discourse competence.

3.6.2 The CEFR and Discourse Competence

Having examined different models of communicative competence and the development of the notion of discourse competence we now turn to the CEFR and the model of discourse competence incorporated in it. The CEFR (Council of Europe, 2001) states that it “does not imply the imposition of one single unified system” (p.7). The document contains only limited references to the sources of the concepts which it draws upon, but it does contain a set of references for each chapter. The lack of detail of sources in the CEFR has been criticised and can make it difficult to clearly identify influences on it but the approach is deliberate. The CEFR is, as it states in its first chapter, an inclusive document intended for use by the widest number of users possible and is therefore intended to be accessible. The document cites Canale and Swain (1980) among its references and presents a view of language that foregrounds the notion that the principle purpose of language learning is communication.

Despite the use of the word 'Framework' in its title, it has been argued by Trim (cited in Fulcher, 2010, p.115) that the CEFR is not a framework, in the sense of being a comprehensive syllabus that can be sampled by course and assessment designers, but a model of language. The point is a valid one and it was not intended that the CEFR would be used directly as a syllabus but as a resource that those involved in the design of courses and assessments could use and apply in their own particular contexts. Page 6 of the CEFR (Council of Europe, 2001) identifies the intended uses of the CEFR and it crucially points out that it is to be used for "the planning of language learning programmes...[and] certification". This use of the CEFR as a planning document is further emphasised by the frequent boxes placed within the document which pose questions to the users of the framework asking them to consider the situation and context within which they are operating. So, if the CEFR is to be considered as a model of language, the issue as to what type of model it is becomes a question.

As a model, the CEFR draws on the notion of competences and identifies these as being "the sum of knowledge and skills that allow a person to perform actions" (2001, p.9). The framework makes a distinction between general competences, those that are not specific to language but are necessary for communication, and communicative language competences which are those concerned with linguistic action. The CEFR divides this latter group into three types; linguistic, sociolinguistic competences and pragmatic competences and it is within the latter that discourse competence is located. Pragmatic competences are defined by the CEFR as being "the functional use of linguistic resources" (2001, p.13) but just as other models of communicative competence have done, the same section of the text stresses the interrelation of the types of competence upon each other. Table 3.1 sets out the components of these competences identified by the CEFR as being concerned with language use.

Table 3.1. *Components of the three communicative language competences as set out by the CEFR (2001, p.108-130).*

Linguistic competences	Socio-linguistic competences	Pragmatic competences
Lexical competence	Linguistic markers of social relations	Discourse competence
Grammatical competence	Politeness conventions	Functional competence
Semantic competence	Expressions of folk wisdom	Design competence
Phonological competence	Register differences	
Orthographic competence	Dialect and accent	
Orthoepic competence		

The CEFR, in order to maintain its methodologically ‘neutral’ position, contains very few in-text citations of its sources, but Canale and Swain’s (1980) article as well as the work of Hymes (1972) is referenced in the bibliography for chapter four, the chapter which sets out the CEFR’s notions of competences. As stated earlier in this chapter, the work of Van Ek (1979) was a precursor to the CEFR and the current document takes language functions as one of the ways in which the ‘can do’ statements are described. All of this underpins the functional approach to communicative competence which underpins the CEFR.

The three competences set out in the pragmatic competences are discourse competence, functional competence and design competence. The CEFR defines these in the following terms (2001, p.123):

- Discourse competence: the way in which messages are “organised, structured and arranged”;

- Functional competence: the way in which messages are “used to perform communicative functions”
- Design competences: the way in which messages are “sequenced according to interactional and transaction schemata”.

The divisions between the discourse, functional and design competences are not altogether clear. Discourse is described in terms of coherence and cohesion in a text (whether spoken or written) which is similar to the original representation of discourse competence in Canale and Swain (1980). However, the framework also includes knowledge of text design or “knowledge of the design conventions in the community e.g. how information is structured in realising the various macrofunctions.” (Council of Europe, 2001, p.123) as part of discourse competence. This inclusion of a more macro-level view of discourse is very close to Swales’s (1991) notions of discourse community and genre and Canale’s (1983) definition of discourse competence. However, the actual macro and micro functions through which discourse competence provides structure to come under the headings of design and functional competences. The macro-functions identified by the CEFR (description, narration, argumentation etc.) seem to be what Smith (2006) would term ‘discourse modes’ or Fairclough (2003, p.69) would call a ‘pre-genre’. So, the relationship between the discourse and functional competences in the CEFR appears to be that discourse competence manages the production of a text according to the speaker’s (author’s) knowledge and experience in a way that is coherent and cohesive and in line with the expectations of the intended readership is managed by discourse competence. However, the actual linguistic forms used to carry out the functions within the text come under functional competence. The CEFR does not set out what comprises design competence but it would appear to be the interface between discourse and functional competence where schemata, coherence cohesion, knowledge of genre and function come together to create texts.

Although both the 1981 Canale and Swain model and the Bachman (1990) model are cited in the CEFR, Swain's 1983 model is not, although the fact that discourse competence is established as a competence in its own right in the CEFR rather than as a subsidiary of grammatical or sociolinguistic competences suggests that Canale's model has been considered. The Celce-Murcia, Dornyei and Thurrell (1995) model as stated above draws on both of Canale's models as well as on the work by Bachman and Palmer (1993). Although the Celce-Murcia *et al*/ model is not cited in the CEFR, its concepts overlap with many of the features of discourse competence described in the CEFR. This model is also useful because its designers were keen to emphasise the interactional nature of the five competences and their role in creating communicative competence.

3.7 The role of discourse competence in determining level in the CEFR

The CEFR not only provides a model of communicative competence which includes discourse competence, it also makes the use of discourse by learners an indicator of proficiency level. Some of those commenting on the CEFR have expressed the view that discourse holds the key to performance in the upper levels in particular (Saville cited in Weir, 2005a). Within the CEFR document itself, discourse is emphasised as being a significant feature, particularly in the higher levels (B2+, C1 and C2).

The lower levels of the CEFR tend to focus on the imparting and receiving of information by learners and the functions around doing this e.g. at B1 a learner can "take messages communicating enquiries, explaining problems etc." (Council of Europe, 2001, p. 34). In terms of cohesion, up until B2, the statements in the CEFR are mainly focussed on clause-level linking e.g. "can link a series of shorter, discrete elements into a connected, linear sequence of points" (Council of Europe, 2001, p. 125). The B2 level though is signalled by the authors as being the start of something different as they state that the B2 level marks "a break with the content so far". In the descriptors for coherence and cohesion there is a

marked shift in the way that the level is described with the can do statement “can use a limited number of cohesive devices to link his/her utterances [*text*] into clear, coherent discourse...” (Council of Europe, 2001, p. 125 *my italics*) as well as “plan what is to be said and the means to say it, considering the effect on the recipient/s” (Council of Europe, 2001, p. 35). All this points towards a shift by learners towards greater discourse awareness and more awareness of the impact of the texts they produce on recipients (Council of Europe, 2001, p. 35).

The B2+ or ‘Strong Vantage’ level is described as being defined by “a new focus on discourse skills” including “argument and social discourse” (Council of Europe, 2000: p.35). Learners at this level are also described as being able to use a limited number of cohesive devices link sentences together smoothly into clear, connected discourse; use a variety of linking words efficiently to mark clearly the relationships between ideas; develop an argument systematically with appropriate highlighting of significant points and relevant supporting detail.” (Council of Europe, 2001, p. 35). The emphasis on social discourse and text design include “knowledge of the design conventions in the community, e.g. how information is structured in realising the various macrofunctions” a concept that can be traced to Swales (1990) and the concept of the discourse community. An examination of the level descriptor for C1 in the CEFR (Council of Europe, 2000: p.24) also suggests that the learner should “use language flexibly and effectively for social, academic and professional purposes” while “the discourse skills characterising the previous band continue to be evident at level C1” (Council of Europe, 2001, p. 36).

There are other points at which the CEFR seems to suggest that learners at the C1 level are more aware of the expectations of readers in terms of the task. In the illustrative scale for general linguistic range (Council of Europe, 2001, p. 110) it is stated that C1 learners “can select an appropriate formulation from a broad range of language”. The use of the

word 'appropriate' is interesting because up until this point the previous levels have talked about a 'sufficient' range for communication of ideas. This descriptor seems to suggest that by the C1 level, learners are able to choose language which is the 'right' formulation for the text and communicative purpose.

The prominence of the role of discourse in the higher levels of the CEFR clearly marks it out as being one of the defining features of these levels. However, as was explored earlier, the CEFR provides very little in terms of the functional exponents which are to contribute to this development of discourse. The illustrative descriptors themselves are problematic often giving little account of what development is expected between levels or how such changes will be developed. Weir (2005a) argues that CEFR needs to be more "comprehensive, coherent and transparent" (see also Shaw and Weir, 2007: p.1) for the purposes of language testing in terms of contextual parameters. Further; if discourse plays such a defining role at higher levels, it is not clear from the CEFR what cognitive demands are placed on learners when producing a text and it is also not clear how this awareness of discourse would develop across levels.

3.8 Discourse competence in writing criteria

In terms of how discourse competence is incorporated into tests of written English, there are two key elements: the task itself and the criteria against which the candidate's script is assessed. The task requires particular discourse through the type of text to be assessed, in this study essays. The task can also specify the intended audience and the purpose of the text (Shaw and Weir, 2007, p.89-90). However, in terms of the evaluation of the responses, it is the criteria which can direct raters to focus on discourse.

Very often writing criteria draw on the terminology from the CEFR. The Cambridge ESOL FCE criteria, which was developed to extend up to the C1 level (Lim, 2012) uses the terms from the CEFR's statements on

coherence and cohesion (Council of Europe, 2001, p. 125) with learners moving from “a limited number” of cohesive devices (B2) to a “variety of linking words” (B2+) which are used “to good effect” at C1 (Lim, 2012, p. 10). The ESB Writing criteria also draw on these same statements from the CEFR in almost the same terms but with little specification as to what this “limited number” or “variety” will consist of. From this perspective discourse competence, in terms of cohesion seems underspecified. Criteria do discuss macrofeatures of texts, in terms of organisation and text type, but the microfeatures do appear to be neglected.

3.9 Conclusion

Successive models have refined and extended the model of communicative competence so that it now includes not only grammatical competence but also contextual factors. Discourse competence has emerged as a category in its own right as “the selection, sequencing, and arrangement of words, structures, sentences and utterances to achieve a unified spoken or written text.” It is also “where the bottom-up lexico-grammatical microlevel intersects with the top-down signals of the macrolevel of communicative intent and sociocultural context to express attitudes and messages and to create texts.” (Celce-Murcia, Dornyei, & Thurrell, 1995, p. 13). This definition of discourse competence suggests that by the examination of microfeatures of discourse as well as through the investigation of what aspects of discourse writers attend to when they are creating texts, in other words, the macrofunctions and communicative intent, it may be possible to explore the cognitive validity argument for a writing task. Such an investigation would use both the products and the processes produced by candidates in order to provide *a posteriori* evidence of validity. The use of verbal reports for the investigation of the processes employed by candidates at different levels to incorporate discourse into their writing has been explored in Chapter Two. In order to investigate the products, that is the texts produced by learners, and to find evidence of the development of discourse awareness what is required is a

feature of discourse which can provide a link between the micro-level coherence and cohesion and the macro-level functions of text purpose and authorial intent. The next chapter turns to the discussion of metadiscourse markers as a feature that can provide such a connection.

Chapter Four: Metadiscourse Markers

4.1 Introduction

The previous chapter explored the concept of discourse competence and the way in which it is central to different models of communicative competence. The chapter also established that discourse competence appears to be a key determiner of candidate performances in the B2+/C1 levels of the CEFR. As stated previously, this study aims to investigate how cognitive validity with regards to discourse competence is demonstrated in the timed essay tasks used in the ESB ESOL International Examinations. The study also aims to examine whether the CEFR's predictions regarding discourse competence at the levels of B2 and C1 are accurate.

This chapter now considers the case for using metadiscourse markers to assess the development of discourse competence in learners. The chapter begins by discussing why this particular feature has been identified before moving on to define metadiscourse in 4.3 and how it represents discourse competence in 4.4. The chapter then focusses on defining metadiscourse markers and then examining different schemes in 4.5 which have been proposed for the analysis of metadiscourse markers. The chapter concludes with a review of the conclusions drawn by previous studies (4.7) which have explored the use of metadiscourse markers. These studies have been drawn on to develop the research questions related to the product strand of the study.

4.2 Features of Discourse

As was referred to in Chapter Three, discourse is a vast area of study with many different approaches within it (Schiffrin, 1994, p. 22). The CEFR

provides a list of features which come under discourse competence (see Figure 4.1 below).

Figure 4.1. *Features of Discourse Competence in the CEFR (Adapted from Council of Europe, 2001, p. 123).*

Discourse competence is the ability of a user/learner to arrange sentences in sequence so as to produce coherent stretches of language. It includes knowledge of and ability to control the ordering of sentences in terms of:

- topic/focus
- given/new
- 'natural' sequencing: e.g. temporal
- cause/effect (invertible)
- ability to structure and manage discourse in terms of :
 - thematic organisation
 - coherence and cohesion
 - logical ordering
 - style and register
 - rhetorical effectiveness
- text design (i.e. how written texts are laid out, signposted and sequenced)

The CEFR's list, while not exhaustive, demonstrates that discourse competence operates on two levels. Firstly there is the micro-level which is the arrangement of sentences to create cohesive stretches of text. However, discourse competence also impacts on the macro-level of text design and how texts are structured. An investigation of discourse competence could focus on various features such as genre and the notion of 'moves' in texts (Swales, 1990; Bhatia, 1993) which are the stages a text goes through in order to achieve its communicative purpose. However, moves tend to represent a macro-approach to text analysis albeit one which can then be explored closer. Such a study would also be

qualitative in design. By contrast an examination of theme and rheme or the way in which new and given information is sequenced in texts would mainly be concerned with the micro-level of a text. In this study, what is being sought is a discourse feature which will allow exploration at both the macro and the micro level. The study also aims to empirically examine the question as to whether there are differences in the way candidates at B2 and C1 demonstrate discourse competence so the exploration must start from evidence within the texts themselves. While moves can offer a way of exploring a text, they are often dependent upon the investigator assigning communicative purpose whereas other features of cohesion such as the use of linkers are more obvious in their functional purposes.

4.3 Metadiscourse

Hyland (Hyland, 2004, p. 109) defines metadiscourse as being “those aspects of the text which explicitly refer to the organisation of the discourse or the writer’s stance towards either its content or the reader”. While Hyland (2005, p. 16) identifies that the basic notion of metadiscourse is conceptually neat, he also states that it is often hard to apply in practice due to uncertainty about where its boundaries lie (1990, p. 188). Schiffrin describes metadiscourse as being “the author’s linguistic and rhetorical manifestation in the text in order to ‘bracket the discourse organization and expressive implications of what is being said’” (1994). Schiffrin’s definition is consistent with the division made by many writers (Vande Kopple, 1985; Hyland, 2004; Swales, 2004; Hyland, 2005; Burneikaite, 2008) between metadiscourse and propositional information. In this division, metadiscourse is concerned with the organisation and stance of the writer while propositional information is “distinct from propositional information, that is information relating to the world beyond the text itself (Halliday, 1994, p. 70).

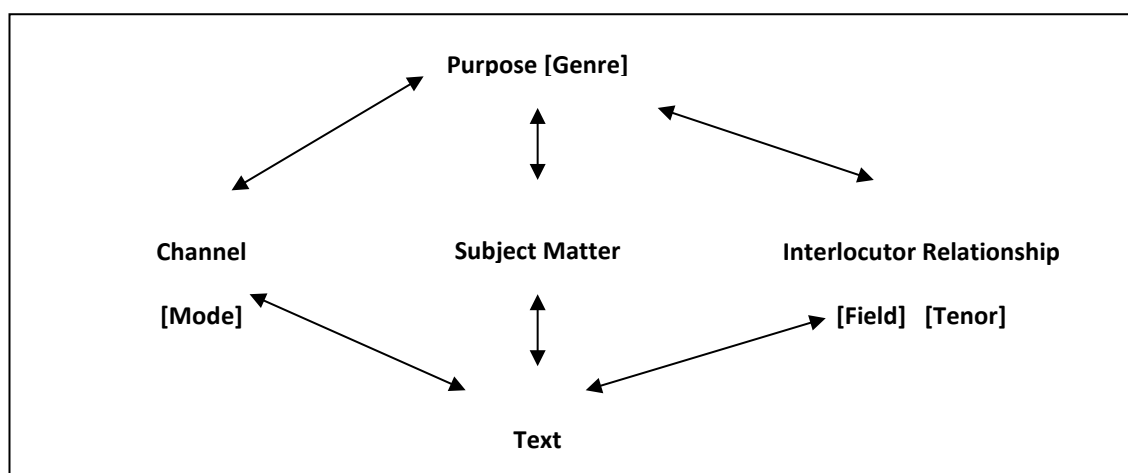
Such a division between propositional information and metadiscourse appears clear but we return to Hyland’s issues about the boundaries of

each. He argues (2005, p. 19) that metadiscourse is not separable from the meaning of the text in that it performs a number of functions which contribute to the effective delivery of the intended message of the text. In other words, a text which is designed to convey a particular communicative intent is organised in such a way as to better facilitate that purpose and the writer's position will be articulated in order to influence the reader towards the writer's intent. Hyland also identifies that writers such as Crismore (1989) have found the division between metadiscourse and propositional information hard to maintain (Hyland, 2005, p. 20-21). Indeed Hyland and Tse (2004, p. 160) point out that removing metadiscourse from a text can change its meaning substantially. They present the case of a scientific article written for a specialised readership being rewritten for a wider audience and while the content is more or less the same, the actual meaning of the text has shifted. Swales (Swales, 1990, p. 188-189) makes the same point when he states that content produced as part of a dissertation has to be rewritten if it is to be adapted into a research article because of the different assumptions about the audience and their level of expertise and background knowledge of the subject.

The point which Hyland & Tse (2004, p. 160-161) are making is that although the distinction between propositional information and metadiscourse is a useful distinction it is not one which can be permanently maintained. The content of a text is inextricably linked to the way in which the message is delivered. Badger and White (2000, p. 155) illustrate the interrelation between different elements of text in Figure 4.2 below. While Field, Tenor and Mode do not entirely overlap with Halliday's (1976) ideational, interpersonal and textual functions, the diagram suggests how these elements interrelate in the process of text composition.

Hyland (2005, p. 37) highlights that metadiscourse can include a wide range of features including non-verbal features including punctuation and forms which might be considered grammatical such as the passive voice or lexis such as adjectives. However, what distinguishes metadiscourse according to Hyland are the functional roles that it carries out within a text.

Figure. 4.2. *A Process Genre Approach to Teaching Writing, (Badger & White, 2000, p.155).*



4.3.1 *The Functional Analysis of Metadiscourse*

As set out in Chapter Three, one of the key tenets of communicative competence is the centrality of context in shaping meaning and language. The notion of this relationship with context is one which Halliday's (1973) distinction between ideational, interpersonal and textual functions helped to shape. Figure 4.2 illustrates this relationship (again, note that field, tenor and mode are not directly interchangeable with ideational, interpersonal and textual functions). Metadiscourse, according to Hyland (2005, p. 24) is best examined through the functions which it performs in a text. However, this necessitates looking at the metadiscourse exponents in context to establish the role that they are being used for in that instance rather than labelling a particular item (e.g. 'should' as an exponent of hedging).

A number of researchers in addition to Hyland have considered metadiscourse from a functional point of view (Lautamatti, 1978; Burneikaite, 2008; Plakans, 2009) and these studies have often viewed metadiscourse as carrying out one of two functions. Firstly, metadiscourse is often viewed as operating at the textual level to provide cohesion between the ideas of the text. These items may be conjunctive and indicate additive, adversarial, causal and temporal relationships in the text (Schiffrin, Tannen & Hamilton, 2001, p.55). Such textual markers “organise propositional information in ways that a perceived audience is likely to find coherent and convincing” (Hyland, 2004, p. 112). This view of textual metadiscourse connects to Halliday and Hassan’s (1976, p. 27) notion of cohesion as being “the means whereby elements that are structurally unrelated to one another are linked together, through the dependence of one on the other for its interpretation.”

However, when Hyland (2004) uses the word ‘convincing’ in his definition he identifies that metadiscourse markers have a role in fulfilling the communicative intention of the writer. Traditionally, this interactional function has been seen as being separate from the textual function. Interpersonal metadiscourse, according to this argument, is used by the writer to indicate their attitude either to the subject matter of the text or to text itself.

Studies have tended to divide metadiscourse into these two categories (Vande Kopple, 1985; Crismore, Martkkanen, & Steffensen, 1993; Intaraprawat & Steffensen, 1995; Burneikaite, 2008) or else to focus exclusively on the textual category (Hawkey and Barker, 2004; Plakans, 2009; Carlsen, 2010) and on explicit cohesive features such as connectives in particular. As in the case of the definition of metadiscourse itself (as highlighted in 4.3) the distinction between interpersonal and textual functions appears clear in that it links back to Halliday’s (1973) model of language functions but the division has been challenged by

Hyland and Tse (2004) and Hyland (2005). Hyland and Tse (2004) argued that even an element of metadiscourse as seemingly uncomplicated as a text connective (e.g. however, therefore) represents the writer's representation of the relationship between one piece of information to another within the text. This link may rely on an assumption of shared knowledge or scholarship with the reader. Essentially, what Hyland and Tse argue is that all metadiscourse is interactional in that it is utilised in order to facilitate the text purpose and the communicative intent of the author upon the reader.

The argument put forward by Hyland and Tse (2004) is convincing but it does pose some questions. The distinction between the interpersonal and textual functions is partly maintained by Hyland (2005, p. 41-50) in the distinction between interpersonal and interactional categories (see sections 4.5 and 4.6) for a detailed discussion of Hyland's 2005 metadiscourse scheme) so perhaps the issue with schemes which do divide textual and interpersonal functions is the overlap between the two functions. However, more importantly for the purposes of this study is the question of how this distinction between the two functions may relate to language learners and the CEFR.

4.3.2 Textual and interpersonal functions and language proficiency

As identified in Chapter Three, the CEFR proposes that discourse competence is a key feature in marking out proficiency at the higher levels of the framework from level B2+ onwards (Council of Europe, 2001, p. 35). Within the B2 level one of the descriptions is that users will begin to "consider the effect on the recipient", a skill which is further developed in the B2+ band. Shaw and Weir (2007, p. 49) in their description of the cognitive demands of the First Certificate in English Examination, a B2 level examination, suggest that at this level there is increasing emphasis on fulfilling the purpose of the task and that candidates are expected to consider their audience in their compositions. It is, according to the

authors, a first step into knowledge transforming rather than knowledge telling (Bereiter & Scardamalia, 1987 see 2.5.1.4) in that thought must be given to the most effective arrangement of the information in the text for the purposes of the text.

The issue here with Hyland's (2005) identification of all metadiscourse as being essentially interpersonal is that while such an observation can be made very strongly for writing produced by skilled writers, there may be doubt when it is applied to the writing of unskilled writers or those learning to write in a second language. Bereiter and Scardamalia's models (1987) suggest that unskilled writers producing text in their native language will knowledge tell, that is produce text in an additive and relatively unplanned way without extensive consideration of the demands of the task or the expectations of the reader. In the case of a language learner, the CEFR's statements in the previous paragraph suggest that discourse competence is very much an emerging system in the B2, B2+ and C1 levels.

Candidates may consider some aspects of how their text will impact on the reader and the communicative purpose but it is not clear that the interpersonal aspects of categories such as connectives, as highlighted by Hyland, will inform the decisions that they make about a text. If anything the literature suggests that learners may be more likely to take a textual view of metadiscourse functions. Burneikaite's (2008, p. 45) comparison of L1 and L2 produced master's theses concluded that there was "significant overuse of text-connectives" and a general underuse of "reader-orientated markers" (e.g. 'you', 'the reader', 'contrast'). Papers by Kennedy, Dudley-Evans, & Thorp (2001) as well as Hawkey and Barker (2004, p. 150) identify a tendency for even C2 level candidates to shoehorn logical connectives into texts (e.g. therefore, furthermore). These studies suggest that it is possible that learners, even at the higher levels, still see metadiscourse as being a tool for the management of their text and its internal structure.

Anecdotal observations by the author of the work of high-level non-native speakers (C1 and above) also suggests that these learners may be unaware of the potentially negative impact of discourse features on their texts, such as repeated use of the word 'however'. The word is simply used to organise what is being said without consideration of the effect on the reader and such writers often respond with surprise when this is pointed out to them. Intaraprawat & Steffensen (1995, p. 266) state that textual functions such as connectives are relatively transparent as concepts to learners and are therefore viewed as a tool for text organisation while interpersonal functions require more insight into reader-writer relationship. They also make the point that logical connectives such as 'however', 'therefore' and so on are also staples of many text books and that learners are therefore more likely to have been exposed to these and consequently they are more likely to have acquired these exponents for use in their writing.

Even if we accept Hyland's claim that all metadiscourse is essentially interpersonal, it is not clear that this is how learners view it and even in the higher levels of the CEFR, from B2 upwards, the awareness of an interpersonal function to metadiscourse must be regarded as an emerging characteristic and any examination of the use of metadiscourse across the levels would need to consider the use of metadiscourse by learners in the light of this.

4.4 Metadiscourse markers as evidence of discourse competence

Thus far in this chapter it has been argued that metadiscourse is a feature of discourse competence worth focussing on in order to evidence increasing learner sophistication in writing because:

- metadiscourse operates on textual and interpersonal levels thereby providing insights as to how a writer is consciously shaping a text as well as taking into account the expectations of the reader;

- metadiscourse provides a link from the macro-levels of text development (i.e. how a text is to be organised to best fulfil its intended purpose) to the micro-levels (how different functions are carried out with linguistic exponents).

The link from the macro-level to the micro-level in metadiscourse is the use of metadiscourse markers in candidate texts. Metadiscourse markers are the linguistic exponents through which the different functions of the text are carried out. Metadiscourse markers lend themselves to investigation because firstly they are evident in the examination scripts produced by learners and therefore represent an *a posteriori* product. Secondly, metadiscourse markers provide a link between discourse competence and the actual linguistic features of a text. McCarthy and Carter (1994, p. 174) identify that the competent use of discourse will necessarily be reflected in the use of linguistic forms because they are realised through these: “linguistic competence cannot be separated from discourse competence”.

However, an investigation of metadiscourse markers as evidence of discourse competence cannot simply be done by ascribing a particular discourse role to particular linguistic items. For example, in Hyland’s 2004 metadiscourse marker scheme ‘or’ is identified as a ‘code-gloss’ (2004, p. 191). In other words, ‘or’ is used to rephrase part of a text in order to ensure that the reader understands it (in much the same way as the chunk ‘in other words’ is used at the start of this sentence). However, to simply count all instances of ‘or’ as being examples of ‘code-gloss’ would be a mistake. Consider the examples below, all taken from the British National Corpus user interface (Davies, 2004).

- a) Each cast member created his **or** her offstage character
- b) He caught us on the wrong foot once **or** twice
- c) the tell-tale signs of brutal, cruel **or** inadequate parenting

- d) They may be intent on impressing each other or the chairman **or** involved in
- e) The job of a grub, a maggot **or** a caterpillar is simply to eat.

As can be seen from these examples, 'or' is used as part of a particular chunk in a and b and in d to connect two clauses suggesting an alternative. Examples c and e are closer to the notion of a code gloss in that they provide additional examples to clarify or define what is being said. The example given here illustrates that the identification of word or chunk as being a metadiscourse marker cannot be done simply by equating form and function. It must instead be carried out by identifying the function that a word is carrying out within a particular text (Hyland, 2005, p. 24-25).

4.5 Metadiscourse Schemes

A number of different schemes for metadiscourse markers exist. Initial attempts to identify metadiscourse markers such as Zellig S. Harris in 1959 (reported in Skulstad, 2004, p.72) identified metadiscourse markers as being those parts of texts which "contain information of only secondary importance". This definition is unsatisfactory as it is unclear how this distinction between primary importance and secondary importance is to be made. Secondly, the interrelation between the different elements of metafunctions (the ideational, interactional and textual) shows that these 'secondary' elements also contribute to the purpose of the text and the act of communication intended by the writer.

Later definitions of metadiscourse markers (such as those by Vandepol, 1985, p. 83; Crismore and Farnsworth, 1990, p. 119; Hyland, 2004, p. 109) are clearer in that they focus on the role that metadiscourse plays as the language used to organise a writer's message as well as allowing for authorial comment on the content in order to achieve the intended communicative purpose of the text. One of the key elements of

metadiscourse markers, as set out by Vande Kopple is that they do not contain any propositional meaning, what Halliday (1973) would consider to be the ideational level of language. According to Vande Kopple's scheme, metadiscourse markers do not convey information about the state of the world, though the interpersonal elements can indicate how an author feels about a particular piece of information (e.g. its 'truth' or validity). Hence Vande Kopple sets out that metadiscourse markers "help our readers organise, classify, interpret, evaluate, and react to such material. Metadiscourse, therefore, is discourse about discourse or communication about communication" (1985, p. 83). However, this distinction between the propositional and metadiscourse is not set in stone. As previously discussed, Hyland and Tse (2004, p. 161), in reviewing Vande Kopple's definition, argue that metadiscourse should not merely be considered a framework for the message, but as contributing to the effectiveness of the message.

There have been various approaches to subdividing metadiscourse markers into different sub-groups within the interpersonal and textual functions proposed by Halliday (1973). Vande Kopple (1985), in a scheme later developed by Mauranen (1993), proposed seven categories: text connectives, code glosses, illocution markers, validity markers, narrators, attitude markers and commentary, along with a division of markers into interpersonal or textual roles. Text connectives are seen as being purely textual in nature in that they "serve to connect building blocks of information to each other (Vande Kopple, 1985, p. 84). However, as Khajavy, Asadpour and Yousefi (2012, p. 149) identify, there is frequent overlap between the proposed categories in Vande Kopple's scheme and the same metadiscourse marker could be argued to be fulfilling several functions simultaneously. This is not an issue unique to the Vande Koppel model as Hyland (2004, p. 109) and Skulstad (2002, p. 72) describe the term 'metadiscourse marker' as being "fuzzy".

A further model of metadiscourse, derived from Vande Kopple (1985) and based on the modifications made by Crismore, Martkkanen and Steffensen (1993) was developed by Burneikaite (2008, p. 39) which proposes the division of metadiscourse elements into three categories: (i) text organising, (ii) participant orientated and (iii) evaluative. However, examination of the model seemed to raise questions regarding the overlap between some of these categories. For example, *mitigation markers* seem to be a combination of hedging, discourse labels and inclusive markers, thereby making it difficult to allocate a marker to a particular category as the reader must effectively second guess the intention of the writer. To some extent, this is an issue with all metadiscourse schemes, and as Hyland points out (2004, p.113) devices may have multiple functions. Analysis of metadiscourse cannot be said to reveal authorial intention, it can only signal relations within a text from which the authorial intention may be inferred. However, the clearer the categories of metadiscourse markers the easier it is for a researcher to make consistent judgements about usage.

4.5.1 Hyland's metadiscourse schemes

Hyland (2004) developed a list of metadiscourse items which consists of textual and interpersonal markers, based on Halliday's (1973) division of functions into the ideational, textual and interpersonal functions. Textual markers signal the language used to organise the text while interpersonal markers manage the social dimensions of the task and allow for commentary on the intended message by the writer.

Hyland's 2004 model is one of the most comprehensive lists available, and divides the markers into *Textual* and *Interpersonal*, although this division must be regarded as essentially for initial analysis purposes given the comments made in Hyland and Tse (2004). Hyland's 2004 list has over 300 discourse markers within it, subdivided into ten categories with a further four types listed under 'Frame Markers'. However, this list was

updated and reconfigured by Hyland (2005, p. 50-54) in order to fit his updated model of metadiscourse based on the view that all metadiscourse is essentially interactional. The differences in the categories is shown in Table 4.1.

Table 4.1. *Hyland's categories of metadiscourse markers (2004 & 2005).*

Hyland, 2004	Based on Hyland (2005, p.50-54)
Textual	Interactive
Code gloss	Code gloss
Endophoric markers	Endophoric markers
Evidentials	Evidentials
Frame markers: <ul style="list-style-type: none"> • Sequencing • Label stages • Announce goals • Topic shifts 	Frame markers: <ul style="list-style-type: none"> • Sequencing • Label stages • Announce goals • Topic shift • Transition markers
Logical connectives	
Interpersonal	Interactional
Attitude markers	Attitude markers
Emphatics (boosters)	Boosters
Person markers	Self-mention
Relational markers	Engagement markers
Hedges	Hedges

The other important contribution made by Hyland's 2005 work was his attempt to establish a clear definition of metadiscourse which would serve as a reliable basis for the coding of metadiscourse markers. The need for consistent and coherent principles in the coding of metadiscourse is essential. As we have established, metadiscourse needs to be separated from the ideational functions of language in order for it to be analysed effectively. Also, as set out in 4.4 above, since the functional exponents

of metadiscourse may also have ideational functions, there must be a clear way of establishing whether a word or phrase is carrying out a metadiscourse role in the text within which it is contained.

Hyland (2005, p. 38) makes three statements about metadiscourse which serve as the principles for the identification of metadiscourse in texts.

- a) Metadiscourse is distinct from propositional aspects of discourse;
- b) Metadiscourse refers to aspects of the text which embody writer-reader interactions;
- c) Metadiscourse refers only to relations which are internal to the discourse.

Propositions a and b have already been considered within this chapter and we have argued that these can be accepted as long as we also take the following provisions in mind. Firstly, the division between propositional and metadiscourse content is not absolute because the same piece of language can carry out both functions simultaneously. Our second proviso is to do with non-native speakers who are learning the language for whom discourse competence is an emerging feature. For these writers, it is not clear that they are necessarily aware of the interaction between reader and writer and the use of metadiscourse may be closer to the traditional view of textual metadiscourse. Both of these observations will have implications for the analysis of metadiscourse markers and this are discussed further in Chapter Five in 5.11.

Proposition c is Hyland's test to determine whether a particular piece of language is carrying out a metadiscourse function. In order to analyse it, the analyst must consider whether the linguistic exponent under examination is representing a state or relationship from the external world (i.e. it has an ideational function) or whether the piece of language is being

used to manage the text or represent the writer's interaction with the reader.

An example of this is the use of the word 'and'. In most metadiscourse schemes, this word would be regarded as a relatively unproblematic linker or logical connector (for example, Crismore, Martkkanen, & Steffensen, 1993, p. 47), Hyland would argue that the word can be internally or externally orientated. Below are two examples of the use of 'and' both taken from the same text (Jones, 2014).

- (a) directed by Ben Wheatley, the British director of such witty, weird and woozily violent films as Kill List, Sightseers **and** A Field In England.
- (b) Everyone has their favourite Doctor **and** my hunch is that Capaldi will one day be viewed as the

In extract (a), 'and' has the function of adding a new activity to a list, therefore the use of the word in the text is not to extend an argument but to simply represent a real-world state of affairs (i.e. it is not an opinion that Ben Wheatly has directed these films). However, in extract (b), the writer used 'and' to add a viewpoint to another argument. Therefore, in an examination of metadiscourse, sentence (a) would be rejected as propositional while (b) would be included in the count. There is more discussion on the examination of 'transitions' below.

4.6 Categories of metadiscourse marker

Hyland's 2005 list of metadiscourse markers has been adopted as the scheme to be used in this study due to it reflecting the interactive and interactional division and being more comprehensive than the Hyland's 2004 list. The following subsections explore each of the categories of metadiscourse marker in order to define them and their functions in writing.

4.6.1 *Interactive metadiscourse markers*

As illustrated in Table 4.1, interactive metadiscourse markers in Hyland's 2005 scheme carry out many of the functions that were previously considered to be textual functions or those elements which "help the writer to signpost the structure of the unfolding text and to signal the structural links between the various parts of the developing argument" (Burneikaite, 2008, p. 39). Hyland's 2005 scheme views this use of metadiscourse to be part of the interpersonal function, in that the writer is shaping the text with regard for the expected reader and their assumed knowledge, background and the text purpose. The interactive elements do nevertheless represent language which is concerned with the organisation of the text and as has been argued, for those who are developing their proficiency in English language, the interpersonal function may not be something of which they are aware overtly.

4.6.1.1 Transition markers

In various schemes this function has been made up of what were called 'logical connectives' or 'text connectives'. Transition markers are exponents which are used in a text to carry out the functions of adding text (e.g. 'and', 'also', 'moreover'), contrasting content (e.g. 'however', 'despite the fact that') or expressing causation (e.g. 'because', 'so', 'as'). Carlsen (2010, p. 193) refers to this group as being 'discourse connectives' or 'connectives'. Fraser (1996) describes connectives as having the function of linking a segment of text with a previous stretch of discourse and identifies that connectives can function at both sentence and paragraph level.

Many writers have considered the logical connective class to have a purely textual function (Vande Kopple, 1985; Crismore, Martkkanen, & Steffensen, 1993) but Hyland and Tse (2004) have argued that this view of metadiscourse markers is problematic. Hyland and Tse illustrate in their article how textual linkers can be seen as elements of the interaction

between the writer and the reader which often extend or develop an argument in order to meet the expectations of the anticipated reader. Concessive linkers (e.g. however, although, nevertheless etc) are noted by Hyland and Tse (p.163) to have a particularly overt interactional function as they effectively orientate the argument to acknowledge the reader's knowledge and to illustrate familiarity with the conventions of the discipline as in the following example (also taken from Hyland and Tse (2004, p. 163):

“The author accepted the shortcomings of the study due to the fact that this was a non-random sample. *Nevertheless*, the study did highlight that ageism is not confined to Western countries alone.”

Hyland (2005, p. 50-51) also suggests that exponents which have been considered to be ‘logical connectives’ may at times be used for propositions or an ideational function and that there is consequently a need to separate ideational uses from interpersonal uses. An example of this is given in 4.5 above using the word ‘and’ from a newspaper article. The work of Martin and Rose (2003, p. 113) identifies four types of conjunction; addition, comparison, time and consequence and then goes on to examine the different roles that these markers can have in texts in terms of ideational (external) and interpersonal (internal) functions.

Hyland draws on the distinctions made by Martin and Rose (see Table 4.2 below) in order to provide a criteria for the distinguishing of metadiscourse functions from propositional information. Figure 4.1 below illustrates the importance of this distinction by analysing a section of text taken from one of the C1 samples from the pilot data (see Chapter Five).

Table 4.2 Hyland's (2005, p.51) representation of internal and external roles of transition markers based on Martin and Rose (2003, p.127).

Relation	External	Internal
Addition	Adding activities including chronological sequencing	Adding arguments
Comparison	Comparing and contrasting events, things & qualities	Comparing and contrasting arguments and evidence
Consequence	Explaining why and how things happen	Drawing conclusions or countering arguments

Figure 4.3. *External (propositional/ideational) and internal (metadiscourse/interpersonal) roles of transition markers.*

To achieve this, they have to develop a certain method of planning the lesson and divide their time efficiently. **(a) Furthermore**, a good teacher needs to be friendly and cool towards students. It is common knowledge that respect cannot be obliged, but gained. **(b) So**, the teacher needs to gain students' respect by trying to understand them **(c) and** caring about them, **(d) as well as** making the lesson seem funny **(e) and** simple. **(f) And**, finally, as far as I am concerned I believe that what makes a teacher good is their love for the job.

In Figure 4.1 'furthermore' carries out the function of adding to the writer's argument (the writer is arguing what they consider to be a 'good' teacher). Therefore, in (a), the **function** is internal because an argument is being added. (b), by contrast has an external function because it is explaining how teachers gain respect and is a consequence of the statement **that** respect is not obliged but given. (c), (d) and (e) are also external in that

they have the function of adding activities to the list of what a teacher has to do in order to earn respect. However, the use of 'and' in (f) is noticeably different. It links into **personal** observations ("as far as I'm concerned...I believe") before adding to the writer's argument of what makes a good teacher.

The close examination of the internal and external roles of transition markers and the functions of exponents is a highly important contribution to the classification of discourse markers. Whereas previous schemes would have automatically included all of the linguistic exponents highlighted in Figure 4.1 as transition markers (or logical connectives), Hyland's 2005 scheme necessitates a more careful consideration of the function being carried out in the text and the contribution that the transition marker is making. It can also inform us as to whether there are differences in the functions to which candidates put particular transition markers.

As a group of words, logical connectives are identified by the CEFR as having a particular role in illustrating increasing proficiency. An indicator of the B1 level is that the learner is able to "write simple connected text" (2001, p. 26) by linking "a series of shorter, discrete, simple elements into a connected, linear sequence of points" (p.29) while learners at the A2 level are described as being able to link words using 'and', 'but' and 'because'. This description is interesting because it is one of the only explicit references to actual linguistic exponents in the whole of the CEFR. Carlsen's study of logical connectives in Norwegian identified that there is indeed extensive use of the simplest structures from this category of metadiscourse markers by lower level learners. Her study also concluded that as proficiency increases learners reduce their reliance on these simpler forms in favour of more complex markers to carry out the same connective functions. Carlsen (2010, p. 204) also found that in addition to using more complex connectives at the B2/C1 level, overall use of

connectives decreased at the C1 level and she hypothesised that this was due to the use of other cohesive devices by the learners. Carlsen's study suggested that the CEFR underestimates the range of connectives used by lower-level learners (2010, p. 203) but also found that accuracy of use as well as range developed as learners increased in proficiency and that higher-level learners had a tendency to use less high-frequency connectors ('lexical teddy bears' as Carlsen (2010, p. 204) terms them following Hasselgren (1994)).

As was stated earlier in this chapter, Burneikaite (2008, p. 43), who was comparing masters dissertations produced by native and non-native speakers, found that the Lithuanian L2 writers had a tendency to overuse logical connectives, a finding consistent both with other studies of L2 writing (Kennedy, Dudley-Evans & Thorp, 2001; Hawkey & Barker, 2004). In addition to attributing the overuse of logical connectives to the L2 writers being overly aware of the structure of the text, Burneikaite suggests that the over-reliance may be the result of the overemphasis of these features in teaching. This explanation has the ring of truth to it in the experience of the author, who has participated in writing standardisation meetings where the overuse of particular linkers (e.g. 'moreover', 'however', 'in addition') are a common point of discussion and which is usually ascribed to over-teaching.

4.6.1.2 Framemarkers

Hyland (2004, p. 112) identifies framemarkers as being those elements of a text which are used to signal transitions between the different parts of the text. These elements often provide information about what is to come next in the text (e.g. 'To conclude') and therefore often have a signposting function to them. In the writing of essays framemarkers are very important since an essay is a continuous piece of flowing text which does not rely on headings for its structure but on in-text signals. As with transition markers, it is important to distinguish the use of markers to organise an

argument from instances where the same items are being used to sequence events in time which would be an external reference rather than an internal one (Hyland, 2005, p. 51).

The Framemarkers category includes several functions: sequencing, label stages, announce goals and topic shifts. Like transitions, these categories fall into the area of textual metadiscourse in other schemes (Crismore, Martkkanen, & Steffensen, 1993; Hyland, 2004) and the markers are clearly used to organise the text. However, in order for their use to be communicatively competent, the use of such markers must be appropriate to the expectations of the genre and the requirements of the task.

4.6.1.3 Code glosses

Code glosses are those parts of the text where the writer either defines what a particular term means or reminds the reader of the definition that is being applied in the text (Hyland, 2004). Code glosses may be signalled by fairly explicit phrase (e.g. 'for example', 'in other words') or more subtly ('or'). Burneikaite (2008, p. 42) found that although code glosses were generally underused by L2 writers when their work was compared to L1 writers, the difference was not significant. Hyland (2004, p. 113) also points out that while code glosses are listed in his 2004 scheme as being part of textual organisation, they have an interpersonal function also in that they represent the writer's view of the extent of the knowledge that the reader possesses and in Hyland's 2005 scheme code glosses are part of the interactive resources.

4.6.1.4 Endophoric markers

Endophoric markers are those which refer the reader to another part of the text (e.g. 'see chapter two'). Hyland (2004, p. 112) suggests that the use of an endophoric presupposes that the reader has access to the complete text. It is expected that the category of endophoric markers would have relatively poor representation in timed-essay writing tasks due the relative

brevity of such texts and the lack of features such as diagrams, charts or Figures.

4.6.1.5 Evidentials

Evidentials are described by Thomas and Hawes (1994, p. 129) in their discussion on reporting verbs as being markers that explicitly signal “the attribution of propositional content to a source outside the author of the article in the current situation”. Hyland (2004, p. 112) however draws a distinction between evidentials whose function is purely for citation and other reporting words and phrases such as attitude markers which are used to convey the author’s stance to the propositional content.

Evidentials are an important category in fields such as academic writing, where it is essential to attribute accurately. However, in timed essay writing under examination conditions where there is no source text to work from, evidentials are likely to be a very under-represented class. Candidates usually do not know what subject area they will be asked to write about in the examination, so the amount of preparation that candidates could do in terms of pre-reading and research is minimal. As a result, candidates are unlikely to be able to draw on specific sources in their writing. However, Hyland (2005, p. 51) makes the point that evidentials could include the use of hearsay or appeals to general knowledge with the use of phrases such as ‘experts believe that/according to experts’ where this is used to build an argument and is not simply a statement about the world at large.

The lack of opportunity for learners to demonstrate skills around managing sources and material from other sources is a serious criticism of timed-essay tasks which do not have a reading-into-writing aspect, particularly if a test was intended for entry onto academic programmes on which the ability to paraphrase and cite correctly will be an essential skill.

4.6.2 *Interactional metadiscourse markers*

In addition to organising the text and constructing an argument, writers use metadiscourse to signal their attitudes to what they are writing. The interactional use of metadiscourse has a strong *reader* orientation with the writer using the resources listed below to try to shape the reader's response. In academic writing for example, the use of hedges potentially anticipates objections to propositions by mitigating them and allowing for alternative viewpoints. With regard to second language learners for whom discourse competence is an emerging feature in the B2 and C1 levels, learners may underuse certain categories such as hedges (Morgan, 2008, p. 275) and overuse categories such as boosters in order to "sell" their ideas. Learners may also over rely on a limited range of features, what Hasselgren (1994) refers to as 'lexical teddy bears'. These features will tend to be high-frequency pieces of language such as modal verbs which the students may have had a great deal of exposure to both in texts and via instruction and may not always be used appropriately in writing.

6.6.2.1 Attitude markers

Attitude markers are used by writers to indicate their affective response to a particular proposition (e.g. surprise, agreement, outrage etc.). These markers carry a certain expectation that the reader will share the view (i.e. if a writer indicates that something is 'surprising' it is to be inferred that the reader will respond in a similar fashion). From this point of view, attitude markers are clearly an attempt to share or direct the responses of the reader and therefore illustrate the writer's expectations of the reader.

4.6.2.2 Hedges and Boosters

The term 'hedging' originates from Lakoff (1972, p. 195). Lakoff argued that the presentation of information by writers is neither absolute truth nor nonsense but true to a particular extent. As a result writers employ hedges in their texts to which are "words or phrases whose job is to make things more or less fuzzy". Hyland (1994, p. 241) points out that hedges

can indicate the degree of confidence which a writer has in the particular statement that they are making in a text and that this is often flagged by the use of a modal verb such as 'may', 'could' or 'might'. In the same article, Hyland argues that hedges are also a crucial feature of the reader-writer relationship as it avoids overstating a particular assertion and thereby appearing overly dogmatic. O'Keeffe, McCarthy and Carter (2007, p. 174) identify that different contexts require different amounts of hedging and in academic writing this is often dictated by the level of surety that a writer has about what they are stating and the rhetorical conventions of the specific genre (Hyland K. , 2004, p. 113).

Morgan (2008, p. 171) identifies that hedges have three distinct functions in texts. They:

- Avoid absolute statements;
- Acknowledge the presence of alternative voices;
- Express caution in anticipation of criticism.

In the same article, Morgan reports the case of an Italian student who failed to achieve a higher grade in her writing because she "failed to navigate the "area between 'yes' and 'no'" Halliday (1985, p. 335)". The failure to use hedges in British academic writing makes the writer appear dogmatic, extreme and reductionist in perspective. However, it is also the case that the use of hedging is a particularly cultural feature and that it is not necessarily used to the same extent in all types of text. Hyland (2004, p. 114-116) identifies that text books belonging to the hard sciences tend to use fewer hedges than text books in business or arts subjects. Morgan also suggests that cultural differences in rhetorical styles could be a source for the underuse of hedges (2008, p. 277) and Burneikaite (2008, p. 42) also reported that hedges (termed 'mitigation markers' in her study) were underused in an L2 corpus compared to an L1 corpus.

Boosters are those words which are used to increase the force of statements (Hyland, 2004, p.87) and express greater certainty in what they are writing. Studies into the use of boosters by non-native speakers have been contradictory at points. Burneikaite (2008, p. 45) identified that non-native speakers under-used boosters in comparison to native speakers and suggested that a lack of confidence, a reluctance to venture opinions and cultural differences could be reasons for this. By contrast, Morgan (2008, p. 275) found that learners had a tendency to overuse boosters. The CEFR does propose that at the higher levels of competence from the B2+ level onwards, learners will carry out “appropriate highlighting of significant points”. The study by Morgan found that while learners underused hedges they had a tendency to overuse boosters. Morgan speculates that this is done in order to better validate the claims that the learners make in their writing (2008, p. 275) and again suggests that this tendency may be due to L1 influence. While that certainly could be the case, it is also true that learners tend to learn a core of high-frequency lexis which can be the lexical equivalent of the sledgehammer to crack a nut (e.g. must, will, always, certainly) when used in academic writing. Again we can see how if discourse competence is an emerging feature, then increased sensitivity to the perlocutionary force of such markers and consideration of how they might impact on the reader should be an element which develops across the higher levels of proficiency.

4.6.2.3 Self mention

This refers to “the degree of explicit author presence in the text” (Hyland, 2005, p. 53) via the use of markers such as ‘I’, ‘me’, ‘we’ and ‘our’. As this particular study is looking into timed essay writing, it is to be anticipated that with the development of discourse competence and increasing awareness of the expectations of readers, learners at the higher levels would reduce the instances of self-mention in order to create a more ‘objective voice’ in the writing. That is not to say that academic writing

does not include self-mention. Investigation by writers such as Hyland (2004) and Swales (2004) have uncovered many instances of self-mention but in terms of essay writing, it is a frequently taught convention that essays do not make use of this category of words.

4.6.2.4 Engagement markers

The final category of metadiscourse markers are devices used in the text to either address readers explicitly (e.g. 'you', 'the reader' etc.) or else which are employed to direct the reader's attention (e.g. 'consider...', 'look at...'). The use of such metadiscourse markers represents an active and directive approach from a writer and a certain amount of authorial power.

4.7 Studies into metadiscourse and second-language learners

Burneikaite (2008, p. 38) states that research into metadiscourse is relatively new and while there are a number of studies which examine the use of metadiscourse markers by authors (such as Mauranen, 1993; Hyland and Tse, 2004; Hyland, 2004; Vergaro, 2005; Burneikaite, 2008; Carlsen, 2010) there do not appear to be studies focussed on the use of metadiscourse markers in the timed expository essay writing. A study by Intaraprawat and Steffensen (1995) which did look at a timed essay task was limited to only 12 participants at levels that approximated to B1 and B2 (Paper-Based TOEFL 593 and 513). Studies also often focus on the difference between L1 and L2 native speakers (Burneikaite, 2008), or else on simply L1 speakers (Hyland, 2004). However, the studies which have looked at non-native speaker use of metadiscourse markers have thrown up some interesting findings.

Burneikaite's (2008) study focussed on the use of metadiscourse markers in MA theses and identified that while there was little overall difference in the amount of metadiscourse markers used between L1 and L2 writers, there were significant variations in different uses of the markers. Other studies such as Bax, Nataksuhara and Waller (Forthcoming) have

suggested that overall there is little or no difference in the amount of metadiscourse markers used by those with different levels of language ability but that differences can be found in the types of metadiscourse markers used (interactional or textual).

Non-native writers were found to overuse text connectives such as (therefore, furthermore etc.), a factor which the author attributes to the learners' being overly concerned with issues of text coherence. Hawkey and Barker's (2004, p. 150) investigation of a corpus of Cambridge ESOL scripts identified a similar feature in the scripts of candidates at C1/C2 in the use of what they term 'link words' (firstly, therefore, furthermore), a feature which was also found in the work of Kennedy, Dudley-Evans and Thorp (2001). Kennedy *et al* also identified that lower-level exam candidates often over-used explicit cohesive devices, which the authors attributed to rote-learning (Hawkey and Barker, 2004, p. 137). While Hawkey and Barker's use of 'link' words appears to encompass two categories of metadiscourse marker in schemes such as Hyland's (2004), the findings are similar in nature and based on a group of students of comparable level. Carlsen's study of metadiscourse markers in the writing of Norwegian learners (2010) further supports these observations in that the study found that there was a predictable overreliance on certain highly-frequent text connectives at the lower levels, items which Carlsen terms "connective teddy bears" (p.203) in that such items represent security, especially in the context of a timed examination. Carlsen also concluded that at lower levels the same connectives may be used for a range of different purposes and that there tended to be more errors in those connectives used to express causation or adversarial relations when compared to those used for adding information. Finally, Carlsen identified that the use of some high-frequency connectives, while very frequent in the lower levels, tended to decrease as proficiency increased, replaced by lower-frequency connectives to carry out the same function. Carlsen's study is particularly useful as, although conducted in Norwegian, the

investigation was linked the CEFR, as in the case of the current study, and aimed at testing the predications made by the CEFR regarding learners' use of such connectives.

Burneikaite's study (2008) identified other trends in the texts produced by the L2 MA students. Certain categories of word were underused, including endophoric markers, evaluative markers (including emphatics, e.g. certainly, clearly, proves) and reader-orientated markers. Burneikaite makes an important cautionary point in the study regarding factors which might influence the results of such studies. "overuse or underuse of metadiscourse in L2 texts is not treated here as a 'deviation' from the norm, but merely as a culture-based peculiarity of interlanguage texts" (2008, p. 45). Clearly culture has an impact on the extent to which an author feels it is necessary to either overtly organise or comment on the text as studies such as Vergaro's (2005) study of British and American For Your Information letters show. However, it is the case that some of the features identified in the studies above do impact on the reader's impression of a text and the writer's level of ability, as indicated by Hawkey and Barker (2004), and it is conceivable that factors such as the over or underuse of particular metadiscourse markers could influence the grading of a paper whether timed or otherwise in academic settings or the way in which a text is interpreted by a reader. Intaraprawat and Steffensen (1995, p. 270) also make the connection between writing and the notion of joining discourse communities. If a writer is intent on being part of an academic discourse community, for example as a student at a university in the UK or US, then effective use of metadiscourse features will be expected in structuring their work but also in maintaining "ethical" and reader consideration through features like hedging commenting on a writer's views.

A second major factor which must be considered is that of the impact of the task on the metadiscourse markers produced. This impact could result

from either the genre of the set task, as in the case of an essay, or from the wording of the set question itself. An example of this latter issue from the pilot study reported in the next chapter which showed that when candidates were given a task asking them to discuss advantages and disadvantages, candidates used a higher level of comparison/contrastive metadiscourse markers than in a task where the discourse had not been so overtly flagged in the question title.

Previous studies into metadiscourse markers have either only considered a single set title (Intaraprawat & Stefensen, 1995; Plakans, 2009) or else have used texts with different rhetorical functions (Carlsen, 2010) or else a single type of text but with different topics such as Burneikaite's study on Master's dissertations (2008). The use of the single task, as in the case of the studies by Intaraprawat and Stefensen and Plakans often means that the number of subjects examined are quite small (the former analysed twelve samples while the latter only considered six samples). By contrast, Carlsen's (2010, p. 197) study looked **at** a large corpus and was seeking to identify general traits in the way connectives (e.g. 'and', 'but', 'because') were used between different levels and she acknowledged that the variety of tasks between the higher and lower levels was likely to result in a difference in the functions of the connectives used. To some extent the variety of tasks used is an issue with any study which aims to examine a range of writing produced in tests across different levels since the types of task which are appropriate at one level may well be entirely unsuitable for a lower level due to the complexity of the response required, or conversely the task may not be sufficiently demanding to elicit the range of language being sought.

Of the few large-scale studies, the one carried out by Bax, Nataksuhara and Waller (Forthcoming), which looked at a total of 900 scripts (300 at each of the levels B2, C1 and C2) used a range of questions and did not limit the rhetorical format of the set question. Like Carlsen's study (2010)

this results in a broad picture of general use within the essay genre but within which results may have been influenced by the balance and range of tasks. Certainly such larger-scale studies can make a better case for their results to be more generalizable while the smaller studies are more like case studies. As such smaller studies can provide insights into the writing of students in particular contexts which could then be used to formulate hypotheses for further study or else act as validation exercises for small-scale language tests.

4.8 Conclusion

The literature review has attempted to set out the theoretical background to the current study. Chapter Two considered models of cognitive phases in writing which will be tested in the process strand of the study. Chapter Three explored the notion of communicative competence as a model of successful language use and the role of discourse competence as an indicator of increasing proficiency, particularly in respect to CEFR levels. In Chapter Four, the case for the use of metadiscourse markers as indicators of this developing discourse competence in the product strand of the study has been considered.

The literature review has shown that while there are studies into cognitive phases used by those writing in a second-language, there are few studies which look at timed-essay writing and the CEFR is unclear about the cognitive phases expected from learners at different levels. The process strand of the study will seek to explore this aspect and see whether the descriptors which are provided in the CEFR are accurate. In terms of the product strand of the study, we have seen that timed-essay writing has not been examined in terms of discourse competence other than in quite small samples and that while the CEFR provides some description, it is unclear how discourse competence might be carried out in terms of linguistic exponents used.

Chapter Five turns sets out the research methods for the pilot study and reports the results and then considers the implications for the main study.

Chapter Five: The Pilot Study – Methods, Results and Implications

5.1 Introduction

As set out in Chapter One, this study has two main questions which it aims to answer, these are:

1. To what extent is **cognitive validity demonstrated** in the cognitive phases that candidates carry out while producing scripts at levels B2 and C1 in the English Speaking Board ESOL International Examinations?
2. What is the role of discourse competence in deciding whether a script is classified as being level B2 and C1 of the Common European Framework of Reference for Languages (CEFR) in candidate scripts from the ESB ESOL International Examinations?

These two aims focus on different aspects of timed essay production. The first aim is concerned with the cognitive phases which candidates in the examinations employ. As Chapter Two has shown, these phases are not directly observable. The use of verbal reports should allow a researcher to make inferences regarding which processes candidates engage. This exploration of the candidates' behaviour follows the suggestion in Shaw and Weir (2007, p. 6) that in order to make a case for validation there should be exploration both of the cognitive phases elicited by an exam task (i.e. checking the processes of composition that the task elicits from candidates) as well as the products. For this reason, the pilot and the main study constitute two strands to the investigation: the 'process' study focussing on the first aim of the study. The second aim, the 'product' strand of the study, focusses on the analysis of the scripts produced by candidates in the tests in order to find evidence of the development of discourse competence as discussed in Chapter Three.

Dornyei (2007, p. 75) emphasises the need for a 'rehearsal' of the research in order to ensure that there is a high level of quality in the final study. He also makes the point that piloting is particularly essential for quantitative data because quantitative research tools are intended to glean particular data. If the tools are poorly designed then they will fail to collect the information required for a particular study. Robson (2002, p. 383) also comments on how a pilot study is an opportunity to face the practical problems of converting the research ideas into reality. A pilot study was therefore carried out as a preliminary of the project. The pilot was run in order to test both the overall approach and the proposed research tools (i.e. the verbal report for the process strand of the study and metadiscourse analysis for the product strand of the study).

Section 5.2 of this chapter describes the data collection methods for the investigation into process including the various tools used. Section 5.3 then outlines how the data from the process strand of the study was analysed. Section 5.4 explores the data collection methods for the product strand of the study (5.4) and the analysis process (5.5). Section 5.6 sets out the results for the process strand of the study and 5.7 the results for the product strand of the study. The results are brought together for discussion in section 5.8. The general implications from the pilot for the main study will be explored in 5.9 before sections that set out the implications of the pilot for the process strand of the main study (5.10) and the product strand (5.11).

5.1.1 The approaches used in the pilot study

Traditionally, research has often been described as being divided between positivistic approaches and a range of approaches which perceive reality as a less-objective phenomenon. The type of data collected for a study was perceived as being dependent upon which of the two approaches a researcher was taking. Quantitative data is information which is essentially numeric in nature and is analysed via the use of statistical

methods (Dornyei, 2007, p. 24). By way of contrast, qualitative data involves the collection of data via open-ended methods such as interviews or observations and the collected data is then analysed in a non-statistical way. This division has long been regarded as an exaggeration of the two positions, the result of what Angouri (2010, p. 30) describes as ‘the paradigm war’ of the 1960s and 1970s, when qualitative researchers attempted to define themselves against the traditions of a hard-line quantitative position.

Angouri (2010, p. 30-31) proposes that researchers should take a more pragmatic line when determining their methodology and base their methods on the focus of the research. Dornyei (2007, p. 166) argues that such an approach should not be ‘an anything goes’ or ad-hoc approach to methodology but underpinned by a consistency in the world view, methodology and data interpretation of the researcher.

The use of mixed methods, that is the use of both quantitative and qualitative methods, has become widely established in educational research. Various writers have argued that the approach used in research is not itself what makes the research more or less reliable or valid. Rather it is the robustness with which the research is conducted and the appropriacy of methods of analysis and the rigor applied. Bond and Fox (2007, p. 17) argue that “both quantitative and qualitative approaches have the same starting point: in observation”. Mixed method approaches are not new. As far back as the end of the nineteen eighties, writers had put forward the view that the apparent dichotomy between quantitative and qualitative research was over-stated and that both approaches could be combined to improve the quality of research (Grotjahn, 1987; Woods, 1992, p. 381). Indeed, the argument can be made that so-called objectivity of research, in the quantitative tradition, is undermined by the subjective construction of its own tools of investigation; being socially constructed phenomena themselves. When designing a quantitative

research tool, a researcher is selecting what they feel to be important and excluding other factors, decisions which are made based on their own experiences, knowledge and context. So, while the results can be seen as being 'objective' they may be the product of the researcher's 'version' of reality.

Dornyei (2007, p. 164) points to two key benefits of adopting a mixed methods approach. The first is the traditional goal of triangulation: the verification of results from different sources. The second benefit is that mixed methods allow a phenomenon to be investigated resulting in a real-world construct: complex, multi-layered and possibly contradictory in some aspects but giving the researcher a fuller understanding of the phenomenon.

Referring back to the aims of this study set out in 5.1, the scope for mixed methods is clearly demonstrable. The investigation of the process of examination writing requires the analysis of data collected from candidates during writing; candidates reporting what they are doing, which must be transcribed and analysed. Features can be identified through qualitative analysis but the data can also be examined quantitatively. For the second strand of the study, the investigation of the products of exam writing requires the gathering of a large number of texts so that particular features can be identified. Effectively this is the construction of a small corpus, a resource which allows for both qualitative and quantitative investigation (O'Keeffe, McCarthy, & Carter, 2007, p. 1). A corpus is "a principled collection of texts" (Sinclair, 1991) which may comprise written or spoken texts, or both as in some corpora such as the British National Corpus.

With the division in the study between how discourse competence manifests itself in 'process' and 'product' it is the case that the process strand is more qualitative in nature while the product strand has a more quantitative focus. Section 5.2 starts by outlining the methods used in the

process strands of the study and the analysis methods (5.3). From section 5.4 the methods for the product strand of the study will be addressed.

5.2 Data collection methods for the process study pilot

As stated above, the pilot study had two research questions, one of which referred to the process strand of the study. This question was:

- To what extent is **cognitive validity demonstrated** the cognitive processes that candidates carry out while producing scripts in the ESB ESOL International All Modes examinations?

While the product strand of the pilot study (see 5.4 onwards) involved a small-scale trial of methods which it had been decided would be used for the main study, the pilot of the process strand was different in nature. The process pilot was intended to guide the selection of an appropriate methodology for the main study. The focus was on how participants would report on their own writing: two methods of reporting were under consideration. The first was the use of verbal reports, whereby participants talked about what they were thinking about and doing while writing. The second method was the use of a written report. In this second method, participants were encouraged to note down in the margins what they had been thinking about and doing.

The reason for the focus on these two methods arose out of concerns over the issues of veridicality and reactivity (See sections, 2.5.2.2 and 2.5.2.3). Reactivity is the potential for the reporting process to change the way participants in such studies behave (Barkaoui, 2011, p. 52). Issues of veridicality are to do with the fact that cognitive phases are being investigated which may be unconscious and therefore unobservable so it is not clear that participants are actually describing the actual cognitive phases that they are going through (Barkaoui, 2011, p. 52). 5.2.1

addresses them and how the process pilot study aimed to explore the issues before going on to describe the instruments used.

5.2.1 Design of the process pilot study

Shaw and Weir (2007, p. 6) propose that the use of verbal reports is one method which can be utilised during investigations into the validation of writing tests. Through the use of such reports the cognitive phases employed by candidates may be inferred. Dornyei (2007, p. 148) states that the use of verbal reports offers “the closest connection between thinking processes and verbal reports are found when participants are asked to verbalise their ongoing thoughts while focussing on a task”. The use of such methods is common in research into writing (Barkaoui, 2011, p. 52). The arguments in favour of such methods are that they allow for insights into mental processes, produce rich data and avoid retrospective interpretation by the subject (Green, 1998, p. 10-11).

As described in 2.5.2.2 and 2.5.2.3, reactivity and veridicality have long been the two issues for which verbal reports have been criticised. Ericsson and Simon (1980, p. 222-226) defended the use of verbal reports and stated that such methods can be used to accurately investigate cognitive phases so long as a number of conditions are observed. These conditions are that:

- the task and the reporting must be interrelated (i.e. be based on the same task);
- participants should be required to describe what they are doing, not account for their behaviour;
- the verbal report must focus on information which was heeded in the task and is still in the working-memory, and not require the participant to draw on the long term memory. Therefore the focus of the report must be on the task immediately worked on.

It was intended that the main study would use non-native speakers who had qualifications at either levels B2 and C1 of the CEFR. The researcher was concerned that, particularly in the case of the participants with a lower level of proficiency in English, the cognitive demands of writing and reporting in a different mode [i.e. spoken] would prove too challenging and result in either impoverished reports or change the process under investigation. As reported in 2.5.2.3, Plakans (2009, p. 567) states that some learners reported being 'helped' by the process of thinking aloud, although she noted no discernible differences in performance. However, Plakan's participants were apparently of a higher proficiency level than the B2 participants which the main study proposed to use.

It was decided to use native-speakers for the process phase of the pilot because they could be asked in detail about the demands of both forms of reporting (written and spoken). It was also felt that putting B2 level candidates through a process that was potentially too difficult for them would not be right. Finally, since the focus was on managing issues of veridicality and reactivity, rather than on the actual results, it was reasoned that native speakers would provide more information about the process.

The two methods, the verbal and written report, are detailed in 5.2.3 below.

5.2.2 Participants

Four participants took part in the process phase of the pilot; two in the written reports and two in the spoken reports. The participants in the process pilot study were all native speakers of English and were from the TESOL programme delivered at UCLan.

5.2.3 Design of the verbal reports

As described in Chapter Two the use of verbal reports to research writing is well-established but many of these studies have focussed on individuals

writing in their native language (Kellogg, 1999). There are fewer studies which have used verbal reports with individuals who are in the process of learning a foreign language. One reason for this is the fact that writing in a second language is undoubtedly more demanding than for native speakers. If second language writers are asked to report on what they are doing while writing this is likely to substantially increase the amount of concurrent cognitive processing required. The lower the level of competence in the target language possessed by the writer, the greater the demands on the working memory made by writing and reporting at the same time are likely to be.

The written reporting approach was based on an activity for teaching writing from Brooks and Grundy (1991, p. 91) in which learners made notes about what they were doing in the margins while they composed a text. It was felt that such an approach, being in the same mode as the skill being assessed, might be a less intrusive alternative to spoken reporting as well as providing the participants with the time and space to record their thoughts more effectively.

Regarding the use of verbal reports for the study, a number of factors were considered to be essential:

1. The participants should not be over-burdened by the requirement to report what they were doing;
2. The participants should have as much time for the task (excluding the reporting) as they would under normal circumstances in the actual test;
3. As far as possible the implementation of the verbal report should avoid or at least reduce reactivity and veridicality.

These three principles extended to the written reporting method as well.

In order to reduce the cognitive load on the participants, it was decided that rather than opting for a concurrent report, the reports (spoken or written) would be carried out retrospectively with the participant writing for five minutes and then being asked to report what they had been doing. MacArthur (2006) suggests that such an approach, as used by researchers like Kellogg (1999), avoids overloading the participants during the writing process while continuing to access what has been attended to in the working memory. The principle is also consistent with what is proposed by Ericsson and Simon (1980, p. 226) in that a retrospective account must be carried out immediately following the task to ensure that the information which an individual has been heeding is still accessible within the working memory.

The second principle of the design of the reporting tasks was that participants were only given forty-five minutes (not including the reporting time) to carry out the essay writing task. According to the design specification for the ESB ESOL International Examinations, if a candidate successfully balances their time in the examination, forty-five minutes is the length of time that they would have to compose their piece of writing. So, the time-aspect of the examination was included in the task in order to replicate as closely as possible operational test conditions. It was hoped that by taking this step reactivity would also be reduced as participants would be performing under very similar conditions to those that apply to candidates in the operational test.

In addition to the timing issue mentioned above a number of other steps were taken to try and reduce reactivity in the verbal and written reporting. The participants were not provided with any resources which they would not have had access to in the actual examination, so there was no recourse to dictionaries or any other materials. Also, the researcher also did not discuss the writing task with the participants, answer questions about how to write the task nor ask any direct questions about how the

task was being approached or anything which might have suggested a way in which to carry out the piece of writing. Furthermore, no additional paper was provided explicitly for planning. Participants were told that they could use margins or the paper on which the question was printed for planning, which mirrors what candidates would have available to them in the actual examination.

The second key criticism often made of verbal reports is the question of whether they are accurate indications of the actual processes which candidates are carrying out in order to complete a task. In order to ensure that the participants in the study were reporting only heeded information, the principles for verbal reports which were set out by Ericsson and Simon (1980) were adhered to in terms of ensuring that the reports, whether written or verbal, related directly to the task which the participant was performing and that the report was elicited directly after the task. Furthermore, each participant was asked only to describe what he or she had been thinking about or doing, not account for why they had taken particular decisions during the process of writing. This follows the suggestion by Ericsson and Simon (2002, p.982) that probes should avoid requesting explicit explanations for actions. As already stated, these same principles also applied to the written reports being trialled in the process pilot.

The design of the reports also followed principles set out in Green (1998, p. 50). She summarises her discussion on the collection of verbal report data with a number of stipulations:

- Good instructions are essential.
- Practice is important.
- Prompting may be necessary if silence occurs (or if a participant stops writing in the case of the written report).

These principles were used as the basis for the development of the data collection tools for the verbal and written reports. The research tools for the written and verbal reports in the pilot study comprised:

- a pre-task;
- an essay task;
- a researcher's script;
- a writing sheet for the candidate and question paper;
- a template for field notes to be used by the researcher.

These tools are now discussed below in sections 5.2.4 to 5.2.7.

5.2.4 The pre-task

The act of verbally reporting what one is doing can be a difficult and a challenging activity for participants even when conducted in their native language. Johnson *et al* (2008, p. 159) identify that one criticism which is often made of verbal protocols is that participants may not be very good at carrying them out and therefore some form of training is important to prepare participants for the main task. Green (1998, p. 16-17) argues that without some training and feedback the data collected from participants may be limited and potentially invalidated.

In order to prepare the participants and give them some practice for the actual reporting task, they were given a short pre-task to write (see Figure 5.1 below).

Figure 5.1. *Training task used with participants in the process pilot study.*

Read the email below and write a short reply offering to help your friend. Say how you will help him/her and suggest a time to do this. Write between 50 - 60 words.

Hi,
I've got some great news for you! I'm moving into a new flat soon. It's much bigger than the one I'm living in now. It's also near the city centre where you live so we'll be neighbours! I've got so many things to do

before I move there in three weeks' time. I don't know how I'll get everything finished!

Please write back soon.

Regards

Sarah

Participants carried out the task and were stopped after every minute to either verbally report on what they had been thinking about or doing or else to write down what they had been thinking about and doing, depending on which of the two reporting methods they would be using in the main task. As recommended by Green (1998, p. 16-17) the participants were given feedback after the pre-task to assist them with the main task.

5.2.5 The essay writing task

All four participants wrote the same essay. This was to remove the essay title as a variable which might affect performance as the intention was to examine the effectiveness of the two reporting methods. The essay title chosen was one which had been set for the C1 participants in the product strand of the pilot study (see 5.4.2).

5.2.6 The researcher's script

A key feature of the verbal protocol is to elicit the thoughts of the participants without the interviewer leading the participants or engaging them in discussion. For these reasons a researcher's script is required in that it standardises the interactions between the interviewer and the participants as far as possible (Green, 1998, p. 11). Green (ibid) notes that different individuals will respond in different ways to the same task but the aim of the researcher must be to ensure that such differences are not a result of the reporting prompts being applied inconsistently.

A standardised script means that participants are all given the same information but also prevents the researcher from having to come up with prompts on the spot. This is an important issue because even variations in

prompts could result in different responses. Green notes that the use of a prompt such as “keep talking” is less intrusive and less likely to encourage interaction with the interviewer than “can you tell me what you are thinking?” (1998, p. 42).

The same script was used for both methods being trialled (the written and the verbal reporting). The frequency of the interruptions for both methods was also the same in that the participants were asked to say or write what they had been thinking about or doing every five minutes.

5.2.7 The writing paper and the question sheet

The actual paper on which the participants were to produce their essays was designed with two purposes. Firstly, it mirrored the appearance of the writing booklet that candidates use in the tests by being lined sheets with a margin. However, the margin size was increased so that the participants doing the written report could make their comments alongside the text that they had produced. This was based on the activity from Brooks and Grundy (1991, p. 91) which had inspired the idea of the written reports. By giving the participants more space in the margin it was also hoped that the commentary could be produced right next to the text to which it referred.

5.2.8 Field notes

The decision to use field notes emerged from conducting the interviews in the pilot study when the researcher identified that not only would it be necessary to probe some of the things which the participants had said in the post-interview but that it was also important to follow up some of the observations made by the researcher during the process of composition. For example, a participant might pause for a long period (sometimes up to two minutes or more) after writing a particular word. Field notes allow the researcher in the post-interview to direct the candidate to the same point in their writing and say “you paused for about two minutes here. Can you remember what you were thinking about or doing?”. Such questions are

consistent with Ericsson and Simon's (1980) principles for verbal protocols because the post-interview was conducted directly after the writing task and the pointing to the actual spot in the writing text where the pause had occurred served as a retrospective prompt. Field notes were not pre-planned in the pilot but the researcher found that they were necessary during the very first pilot interview after which they were used throughout the pilot.

5.2.9 Procedures used for data collection

The participants for the verbal and written reports were interviewed individually. The researcher began by explaining what would happen using the information sheet which had been provided to the participants in advance of the interview and then the signed consent form was collected.

The pre-task was carried out (the e-mail task) and any resulting questions regarding the verbal or written report process were answered. When the participant was ready to start the main task then the digital voice recorder was turned on.

The candidate carried out the writing task and then a short interview was conducted based on the pre-planned questions and the observations collected by the researcher using the field notes. Once the interview was concluded the data-recorder was switched off.

5.2.10 Ethical considerations

In the pilot study, all four participants were UK students on the TESOL undergraduate degree programme. All four volunteered to take part and were provided with information sheets 24 hours before the study and signed consent forms. Names were removed from the transcripts.

5.2.11 Transcription

In order to be able to examine the information gained from the verbal protocols, the recordings had to be transcribed. The transcription of verbal protocol data is an important consideration since it is this report which forms the basis for the analysis. The process is not without problems. Kvale (1996, p. 145) points out that the actual term *transcription* means changing from one state to another so the act of turning spoken text into written text will inevitably mean that some aspects are lost or changed. The loss of information is due to the non-verbal information that face-to-face interviews include and which cannot be captured with ease in a written description. Changes to the data are sometimes due to the actual process of transcription. Richards (2003, p. 202-205) gives an example of two researchers who independently transcribed the same meeting and displayed a number of significant variations in their resulting transcripts thereby illustrating how different transcribers can inadvertently represent their data in different ways.

Green (1998, p. 52) emphasises the need for the transcription to be faithful to the original recording in that it should not represent a 'cleaning up' or a 'tidying' of the data, but should report exactly what was said and in what manner. However, she also suggests that other sources of data can be utilised in order to clarify what is meant at different points. During the reporting (see above), the researcher kept field notes and these field notes have were later typed out.

Another point which Green (p.51) makes regarding the faithfulness of transcription is the inclusion of paralinguistic features such as pauses and laughter and the importance of including errors, mistakes and slips made during speech. Since the pilot study used native speakers, language errors were not particularly an issue although features of speech such as false starts, broken sentences and slips were still in evidence and these were retained.

5.3 Analysis of the scripts for the process strand of the pilot study

As set out earlier in this chapter, the purpose of the process phase of the pilot study was to determine which of the two methods, the verbal or written report, was most effective in eliciting cognitive phases from the participants. The pilot study also sought to find out whether one or other of the methods was more effective in reducing reactivity or veridicality. From this point of view it was the post-interview comments on the method which the researcher was particularly interested in. However, a scheme was also needed to analyse the data which the verbal reports had elicited from the native speaker participants.

It was decided to use an *a priori* scheme to code the verbal protocols and Field's descriptions of the cognitive phases (2004, p. 329-331) were used for this purpose. The use of pre-existing categories for coding has been criticised by some writers. Green (1998, p. 70) suggests that the use of such pre-existing schemes have a tendency to be geared towards testing a particular hypothesis and may result in data being excluded which might otherwise be considered. Brown and Rodgers (2002, p. 65-66) point out that that the use of a pre-existing coding scheme may not be suitable for the analysis of the actual data collected for a study. They do however suggest that the use of a pre-existing coding system will already have some level of acceptance in the academic community and that the data can be more readily compared to sets of data from other studies using the same coding system. Crabtree and Miller (1999, p. 164) advocate what they term the 'template organising style', that is using an established code as a basis for investigation where a well-established theory is being tested.

While the notion of carrying out a qualitative analysis with a preconceived approach may seem contradictory, Dornyei (2007, p. 253-254) makes the argument that only very rarely to researchers come to their data without

preconceptions, either developed during the design of the study or through the pre-coding. It was also the case that the main focus of the verbal reports was the question of which method had worked best in reducing the cognitive load on the participants and the possibility of reactivity and veridicality.

Once the recordings had been transcribed, an analysis was carried out on the transcripts from the tasks and interviews. The researcher highlighted each instance where the participant said or wrote a comment that appeared to fit into one of the processes identified in Field (2004).

In order to provide inter-rater reliability, a colleague was also standardised as a rater and asked to check the categorisation of the data from the two types of reports (verbal and written). The researcher was then able to count the instances of implied cognitive processes and compare them between reports as well as looking at the comments made by the participants about the methods of reporting during the post-interviews.

Section 5.6 of this chapter sets out the results from the process strand of the pilot study and these results are discussed in 5.8. The next sections coming up (5.4 and 5.5) consider the data collection methods for the product strand of the pilot study.

5.4 Data collection product

As discussed in 5.2, the pilot of the process study was intended primarily to determine whether a written report would provide richer data and better avoid the problems of reactivity and veridicality. The data collection for the product strand of the project was however intended as a small-scale version of the main study. The following sections discuss the research methods used in the pilot study, starting with the design of the study (5.4.1), the scripts used (5.4.3), how the scripts were rated for inclusion in

the study (5.4.5) and transcribed (5.4.6), the ethical issues involved (5.4.7) and how the scripts were analysed (5.5).

5.4.1 Design of the product pilot study

As set out in Chapter Four, metadiscourse was selected for investigation into discourse competence in this study for the following reasons:

- metadiscourse has interactive and interactional functions. The interactive functions include decisions regarding text organisation on the part of the writer. Interactional functions are those by which the writer seeks to intrude into the text and engage with the reader (Hyland, 2005, p. 218-224). Metadiscourse therefore may offer insights into how a writer is consciously shaping a text and taking into account the expectations of the reader;
- metadiscourse provides a link from the macro-levels of text development (i.e. how a text is to be organised to best fulfil its intended purpose) to the micro-level of the linguistic exponents used. This suggests that metadiscourse may indicate how candidates manage the interaction with the writer (in terms of the interactive and interactional functions) through their choice of linguistic exponents.

As has been discussed in the Chapters Three and Four, discourse and metadiscourse have often been viewed as difficult areas to pin down or define. There are many different aspects of discourse which could be examined in a piece of writing. For this study, the researcher believed that metadiscourse markers would offer up a potential source of empirical evidence in that text analysis software could be utilised to count the instances and thereby provide a reliable basis for investigation. As Levon (2010, p. 68) points out, in order for quantitative analysis to be carried out the object under investigation must itself be capable of being counted and metadiscourse markers allow such a method to be used.

The use of data gleaned from candidate performances in order to examine the validity argument of an examination is an example of what Weir (2005) terms *a posteriori* validation in that it is an approach which uses evidence generated by the test and so only becomes available after the test has been administered. Weir argues that this type of investigation is often concerned with aspects such as the criterion, content or construct validity of an assessment. In the case of this study the intention was to search for metadiscourse markers in order to examine the issue of whether the ESB ESOL International English Languages tests show a change in the awareness and use of discourse by candidates at B2 and C1. The study also aims to see whether the changes in discourse use predicated by the CEFR were evident in the timed writing tasks under investigation.

The texts collected were hand-written, just as the operational examination is. This is because the composition process may alter in terms of the extent of revision that word-processing candidates might carry out when compared to the more limited opportunities for candidates who are producing their text longhand. Studies such as those by Weir, O'Sullivan, Yan and Bax (2005) suggest that producing texts electronically might not have **affected** the grade awarded but because the examination under investigation in the study is administered longhand, collecting the texts electronically might have undermined the validity of the study. As a result, the texts were transcribed (see 5.4.6).

In order to examine the role of discourse markers in candidate writing it was necessary to construct a small corpus of samples of candidate writing at the different levels under investigation. Texts were transcribed and stored on computers for easy access. Tools for the analysis of corpus information such as *Wordsmith* or online concordancers such as those on *The Compleat Lexical Tutor* website are often used but non-specialist word-processing programmes such as Microsoft's Word, although much

more limited, also allow users quickly to search for and locate specific information.

The practice of using small-scale corpora in order to examine linguistic features is common in EAP and ESP settings (Carmen Campoy-Cubillio, Belles-Fortuno, & Lluisa Gea-Valor, 2010). The method is also consistent with the intention to carry out a quantitative analysis of the products of candidate writing. Baker (2010, p. 94) states that the use of corpora in linguistic investigations “is firmly rooted in empirical, inductive forms of analysis” in that it uses real-world data and is based around the amassing of evidence from the texts which make up the corpus.

Luzon, Campoy, Sanchez and Salazar (2007) point out that a key feature in the compilation of a corpus is the purpose and the principles that underpin its construction. While the sources cited above are mainly concerned with the development of teaching materials, the same principles for corpora development extend to building a corpus for the investigation of candidate performance in examinations. Cobb (2003, p. 394) identifies that every corpus must have its own rules for the inclusion of materials. It is these rules which will determine how useful a corpus will be in terms of its output. These rules comprise a number of key questions:

- i. How large does this corpus need to be?
- ii. What types of text have been entered into the corpus?
- iii. By whom have these texts been produced?
- iv. How have these texts been ‘approved’?
- v. How have the texts been transcribed and coded?

Each of the above questions requires consideration in order to demonstrate that a corpus is suitable for the purpose to which it is being put. The following sections (5.4.2 through to 5.4.6) use Cobb’s questions

as the basis to identify the key decisions which were taken in the assembling of the corpus.

5.4. Corpus size

Baker (2010, p. 95) states that there are no clear rules regarding how large a corpus must be. What does matter in the construction is how the corpus is intended to be used and what information it is expected to yield. In terms of the number of words the number required ranges from 100,000 to half a million with Biber (1993, p. 244) suggesting that a million words would be sufficient for a study of the grammatical features of texts. However, McEnery and Wilson (1996, p. 22) state that the number must be “maximally representative of the sample it represents”. Baker (2010, p. 96) adds to this by suggesting that a governing factor must be the variety of text types that make up the corpora.

For the pilot study the decision was taken to keep the corpus small in size with just four scripts at B2, B2+ and C1. The B2+ level was included as the corpus was small in size and four scripts were identified as being B2+ in terms of proficiency. The scripts were selected from the learners described in 5.4.4.

5.4.3 Text types in the corpus

Two writing tasks were selected from past ESB papers at levels B2 and C1. Both tasks and the rationales are reported in Table 5.1 below.

Table 5.1. *Essay tasks used in the pilot.*

Task	Level of Task	Title
A	C1	“I like a teacher who gives you something to take home to think about besides homework.” Write an essay discussing what qualities you believe a good teacher has and how a good teacher influences you in your studies.

Word limit: 300 – 350 words		
B	B2	<p>Many teachers and experts in education feel that the school holidays are too long and suggest that children need to spend more time in school.</p> <p>Write an essay discussing the advantages and disadvantages of having long school holidays.</p> <p>Word limit: 180 – 200 words</p>

5.4.4 *The source of the texts*

The writing samples were collected from three groups (see Table 5.2). Non-Native Speakers at B2 and C1 were both assigned Task A and Task B to carry out. Task A was the designated C1 level task and Task B was the designated B2 Task. 14 scripts were initially rejected as being far too low for analysis in that raters identified them as being level B1. Four scripts were rejected as permission for use had not been given by the candidate and one candidate only completed one of the tasks so their script was also rejected.

The non-native speaker candidates were all students in a private language school in Greece who were preparing to take either the B2 or the C1 examination. All of the candidates declared the results of the last examination that they had passed (including the date). The candidates were aged between 15 and 21 years old and were native speakers of Greek. Fifty-six per cent of the candidates were female.

The native speaker candidates were all undergraduates aged 18 – 24 in the first semester of their degree (the majority were studying TESOL with a modern language).

Table 5.2. *Scripts collected for analysis.*

Group	Total Sample scripts	# of texts removed	Tasks attempted	Final # of scripts in corpus
Non-Native Speakers B2 (NNS B2)	94	19*	A & B	4
Non-Native Speakers C1 (NNS C1)			A & B	4
Native Speakers (NS)	25	0	A	4

** one candidate only wrote one of the scripts*

5.4.5 Approval of the texts

Scripts were put into a task A and task B pack, randomised and passed to three raters for grading against CEFR scales. Raters graded each script as being either below B2, B2, B2+, C1 or C1 or above. The scripts were entered into a computer and analysed against Hyland's metadiscourse markers list (2004, p. 190-193) using textinspector software (Bax, 2011).

This software was used as it was able to identify metadiscourse markers from Hyland's 2004 scheme but also allowed for lexical items to be re-classified by category, or removed from the count following visual inspection of the use of the item in context.

Only four scripts at each level were put forward for analysis for this project, due to the limited number of scripts designated as being C1 in level following rating. However, a set of scripts were also designated as being B2+ and were included in the analysis as this level is arguably where the CEFR identifies the increasing importance of discourse (Council of Europe, 2001, p. 35).

5.4.6 Transcription and coding

As discussed in 5.4.1, a feature of the examination under investigation is that it is a handwritten paper and during the development of the study it was decided that this factor would have to be maintained. There was also the issue of practicality since the examinations in the main study would be administered in Greek public schools where there is limited access to computers and there would not be enough to test the numbers of candidates at B2 in particular.

However, all of the samples collected were presented electronically to the raters in order to minimise the impact of handwriting as a factor in the evaluation of the scripts. Brown (2004, p. 117) has commented that the legibility of a script is known to influence rater decisions although as Shaw and Weir (2007, p. 177) point out, there has been little investigation into the exact impact. The authors do cite a number of studies which suggest that tidy handwriting could speed up marking speed and thereby reduce the strain on the reader and so create a more favourable impression. A recent unpublished internal review of writing criteria at UCLan and ESB (2014) asked the participants (test developers and raters working in four independent groups) to take various features of writing (such as cohesion, task, grammatical accuracy etc) and to consider how they could be combined into writing criteria in different ways. While two groups chose to omit handwriting from their writing scales, suggesting that legibility was instead a factor to be considered in the ground rules of the assessment, one group felt that it was a factor which could be assessed as part of a scheme. This example illustrates that for some raters, handwriting is a proficiency criteria so it was decided that for the study this variable would be removed by presenting all the scripts in typed form.

Texts were typed up by the researcher for entry into the corpus. A number of editorial decisions were taken during this process.

- The researcher only typed up what the candidate had clearly intended to be submitted. Any text which was crossed out or erased was not included.
- A second editorial decision was that no spelling errors or grammatical errors would be corrected. Such errors had to be carefully monitored during data entry as often the autocorrect function of Word (Microsoft, 2013) would attempt to revise these errors. When a piece of text was illegible, it was initially shown to a colleague who had experience in marking second language writing and if it could not be agreed what the word was it was replaced with the word 'illegible' in square brackets and highlighted.
- The only additions made to the text were where a text contained a mis-spelt word or phrase that had the potential to function as a metadiscourse marker. A correction was provided directly after the word in square brackets, in bold. All such corrections were removed before the texts went out for independent rating.
- Finally, texts were stored as 'plain text' files since, as O'Keeffe, McCarthy and Carter (2007, p. 8) point out, this format is the most versatile for the purpose of analysis by different software.

5.4.5 Ethical issues

All participants in the product strand of the pilot study were provided with an information sheet about the study and completed a consent form. Where a participant did not complete the consent form (whether due to unwillingness to participate or forgetting to tick the box) their data was not included. Table 5.2 includes the Figures for those scripts which were excluded due to lack of consent.

5.5 Analysis of the product data

The data for the product strand of the pilot was analysed using *textinspector*, a programme developed by Bax (2011) which automatically identifies and classifies metadiscourse markers using Hyland's 2004

categories (p. 190-193). However, the software can only detect the presence of words designated to have metadiscourse functions. Therefore, the texts then had to be examined by a human rater to confirm whether each word identified by the software was functioning in context as a metadiscourse marker. This process of review also allowed the researcher to identify other words in the text which were not part of Hyland's list but which did function as metadiscourse markers.

The analysis process resulted in a count of discourse markers by category for each script that was analysed. This data was then examined to look at the different functions of metadiscourse markers within categories (e.g. whether logical connectives are being used to add to a clause, to introduce a contrast or to signal cause and effect).

In terms of statistical measures, the small number of samples at each level (4) meant that there were not enough candidates for correlational analysis as Dornyei (2007, p. 99-100) suggests that at least 30 participants are needed for this or for statistical significance to be calculated. As a result, only descriptive statistics were used for the analysis of the data: a limitation on the pilot study (see 5.11 for more on the limitations of the product strand of the pilot study).

5.6 Results from the process pilot

Both the written and verbal reporting methods employed in the process strand of the pilot study demonstrated evidence of reflexivity on the part of the participants, particularly with regard to the written reports, with participants A and B commenting that the process had influenced their composition process.

Table 5.3 (below) records the instances of each of the cognitive processes reported by the subjects. Although there does not appear to be much difference in terms of the number of instances of processes, the data

elicited from the spoken reports was far richer in terms of the comments made during the actual writing process, whereas the written reports tended to require more detailed explanation retrospectively in the post-interview. The post-interview could still be seen as being valid, in terms of Ericsson and Simon's requirement (1980) that it be held while the heeded information was still in the working memory.

Table 5.3. *Instances of cognitive processes elicited from participants.*

Subject	Written Reports				Spoken Reports			
	A	A %age	B	B %age	C	C %age	D	D %age
Macroplanning	5	8.77	9	29.03	10	17.54	5	11.36
Organisation	8	14.04	2	6.45	6	10.53	8	18.18
Microplanning	15	26.32	7	22.58	3	5.26	20	45.45
Translation	15	26.32	5	16.13	11	19.30	5	11.36
Monitoring	13	22.81	5	16.13	19	33.33	6	13.64
Revising	1	1.75	3	9.68	8	14.04	0	0.00
TOTALS	57		31		57		44	

5.7 Results from the product pilot

The results from the product strand of the pilot study are set out in Tables 5.4, 5.5 and 5.6 below. Each of Hyland's 2004 categories of metadiscourse markers are set out below. Table 5.4 shows the means for tokens in the scripts and the type-token ratios for each group.

Table 5.4. *Overview of texts analysed.*

		TASK B (B2 Task)			TASK A (C1 Task)	
		# Texts	Tokens (Mean)	Type/ Token Ratio (Mean)	Tokens (Mean)	Type/ Token Ratio (Mean)
Candidates						
B2	B2	4	326	0.53	311	0.47
B2+	B2+	4	265	0.52	349	0.47
C1	C1	4	263	0.54	429	0.49
NS	NS	4	NA	NA	464.5	0.45

Table 5.5 and 5.6 present the numbers of metadiscourse markers according to Hyland's categories (2004, p. 190-193).

Table 5.5. *Discourse markers by level (average % of text) for Task A (C1 Level Task).*

Candidate's Level	Announce Goals	Attitude Markers	Code Gloss	Emphatics	Endophorics	Evidentials	Hedges	Label Stages	Logical Connectives	Person Markers	Relational Markers	Sequencing	Topic Shifts
B2	0.08	0.80	0.48	2.41	0.00	0.24	1.77	0.08	5.06	0.80	2.57	0.00	0.00
B2+	0.00	0.50	0.21	2.15	0.00	0.43	0.86	0.14	6.30	2.79	2.44	0.29	0.00
C1	0.06	0.52	0.41	2.56	0.00	0.17	0.99	0.17	6.47	0.76	2.97	0.23	0.06
NS	0.00	0.59	0.32	1.83	0.00	0.16	1.78	0.11	3.71	0.27	0.38	0.16	0.00

Table 5.6. *Discourse markers by level (average % of text) for Task B (B2 Level Task).*

Candidate's Level	Announce Goals	Attitude Markers	Code Gloss	Emphatics	Endophorics	Evidentials	Hedges	Label Stages	Logical Connectives	Person Markers	Relational Markers	Sequencing	Topic Shifts
B2	0.08	0.54	0.46	1.76	0.00	0.23	2.07	0.31	6.67	0.77	1.23	0.08	0.00
B2+	0.00	0.19	0.47	0.66	0.00	0.28	0.94	0.28	6.13	0.19	1.89	0.66	0.00
C1	0.00	0.10	0.38	1.62	0.00	0.38	1.24	0.19	6.46	0.10	2.38	0.48	0.10

5.8 Conclusions and discussion

5.8.1 Discussion of results from the process strand of the pilot study

One of the key observations from the pilot spoken and written reports was the extent to which both methods were retrospective rather than completely concurrent in that the tasks required participants to report on what they had just been thinking about and doing in the five minutes before the intervention. Feedback from the participants in the pilot suggested that when they were required to orally report what they had been doing, the participants had less opportunity to filter what they said resulting in richer data and less opportunity for the participants to plan their comments. This seemed to indicate that the verbal report was less open to influence from the long-term memory and afforded the participants less opportunity to plan what they thought they *should* be saying. This issue was a problem with the written report method adopted in that the subjects had time to filter what they wrote and did so. Extracts 5.1 illustrate this:

Extract 5.1

(a)

I think writing it down was easier, erm, because it gave me the time to sit and think whereas if I'm telling you what I was thinking, I probably would have felt the pressure to keep talking, rather than to pause and think what I was saying.

(Participant A)

(b)

[writing] gave me time to think...erm, it was distracting when I stopped about what I was thinking but then it gave me chance to go over everything already and I even looked back at what I wrote, about what I was thinking and thought ah! That could be used. So, it worked out and it helped because it gave me more information.

(Participant B)

This suggested that there was a strong possibility of veridicality affecting the results. For this reason, it was decided to use the verbal report in the main study.

Observations made by the researcher while the subjects were writing, functioned well as a method of stimulated-recall which could be used in the post-interviews.

In summary, the pilot study concluded that the use of a think-aloud method, administered at regular instances during the writing process, coupled with a stimulated recall post-interview task is the more effective of the two methods for eliciting the processes candidates undertake. The method appears to provide richer data during the actual process and appears to reduce the impact of veridicality.

5.8.2 Discussion of results from the product strand of the study

As might have been expected, B2 candidates tended to use more attitude markers and person markers across both tasks (in the latter if one excludes the one B2+ candidate who used 30 instances of *me* or *mine* in Task A). Person markers generally reduced in the higher levels, suggesting growing awareness of the essay genre and the stylistic requirement to avoid use of the first person. Similarly, there was less reliance in the higher levels on modals for emphatics and hedging than at the B2 level.

With regard to use of logical connectives, Kennedy, Dudley-Evans and Thorp (2001) identified that even at the C1 and C2 levels, candidates have the tendency to “learn a set of link words or phrases and force them into their writing”. The data from the current study suggests that the C1 candidates used more of this type of language compared to the B2 candidates. Native speakers used far fewer logical connectives. It seems possible that native speakers organise their texts and the relationships

between ideas utilising other features such as theme and rheme, position in the text and implied relationships.

Sequencing markers (e.g. in conclusion, to sum up) had a low count overall but the B2 candidates used very few (and none at all for the Task A), perhaps because they have less text-level awareness.

Hedging increased between B2+ candidates and C1 candidates, while the native speakers provided considerably more instances of this feature. The B2 candidates had very high scores for hedging, largely due to their heavy use of modals. However, it was felt that there may have been issues with the way metadiscourse markers had been identified in the pilot study in that Hyland's 2004 scheme had been followed and it lacked the later more stringent tests for metadiscourse that his 2005 scheme included.

In summary, the results from the pilot study suggested that there were key differences in the way that candidates at different levels deployed metadiscourse markers in timed essay writing. These differences included not only the number used but also the range of forms used.

5.9 Limitations from the study and implications for the main study

As Dornyei observes, a pilot study allows for a 'rehearsal' of the research (2007, p. 75) in order to ensure that potential problems which might impact on the main study are identified and mediated. Pilots also provide an opportunity to trial the tools which are to be used on the main study. In these respects, the pilot study reported in this chapter worked well as a number of problems were identified and dealt with. The following section starts by identifying common issues which affected both strands of the pilot before discussing particular limitations and implications for the process and product strands of the main study.

5.9.1 Research questions

The two research questions used for the study in the pilot were too broad and were not sufficiently focussed to guide the main investigation. As a result, five research questions were derived from the two aims. These research questions were focussed on the two strands of the study with the aim of ensuring a more focussed main study. The resulting research questions were:

Aim 1: investigation of processes:

1. What cognitive phases do candidates at levels B2 and C1 appear to employ when composing timed essay tasks?
2. To what extent does the timed essay question format used in the ESB ESOL International Examinations elicit the cognitive phases that models predict at levels B2 and C1?

Aim 2: Investigation of products:

3. Is there a difference in the quantity of metadiscourse markers used by candidates of the ESB International ESOL Examinations at levels B2 and C1 of the CEFR?
4. What are the functions of the metadiscourse markers used by candidates at level B2 and C1?
5. To what extent do the metadiscourse functions employed by candidates at levels B2 and C1 match the predictions made by the CEFR regarding the development of discourse competence in learners at these levels?

The product strand of the study also had a number of hypotheses derived from the research questions to inform the quantitative investigation and these are detailed in 6.6.3.

5.9.2 Selection of the writing tasks

A second serious problem with the data collection in both strands of the pilot was the choice of the writing tasks (see Table 5.1 above). Although the nature of the tasks had been considered, there were two substantial problems. First of all, the two tasks did not have the same rhetorical pattern (i.e. advantage/disadvantage) so it was feasible that any resulting differences in the use of metadiscourse markers might have resulted from the different styles of the questions. Secondly, the C1 question was an opinion essay which meant that it tended to encourage knowledge telling rather than knowledge transforming (Bereiter & Scardamalia, 1987) in that it was feasible for writers to simply keep adding ideas in a list rather than having to develop a particular line of argument. This C1 task was also used with the participants in the process strand of the study so again this might have impacted on how they chose to carry out the task and not have required the level of complexity that a B2-level task should demonstrate.

It was decided that for the main study the writing tasks would have to be approved by a raters for the level and that any tasks used would have to follow the same rhetorical pattern in order to reduce this factor as a variable in the study.

5.10 Limitations of the process strand of the pilot and implications for the main study

The process strand of the pilot had used only four participants and it was felt that the main study would have to be carried out with larger numbers. In addition, the pilot had not actually used second language learners at the B2 or C1 levels. This necessitated a further trialling of the verbal report method and the tools for data collection with learners at these two levels before the main study was carried out (see 6.3.2). As reported in 6.3.2, a further trial of the verbal report tool was conducted with non-native speakers at levels B2 and C1 to ensure that the tool was effective. It should be acknowledged that the written reports had only been trialled

with native speakers but it was felt that in addition to potentially causing problems of veridicality and reactivity, further additional writing would place extra demands on participants in that it would be extremely time-consuming.

The main concern of the process strand of the pilot study had been to determine the reporting method to be used (written or verbal) in the main study. As a result of this the analysis of the resulting data from the reports was very limited. It was decided for the main study that a number of changes would have to be made. First of all, while it was intended that Field's (2004) phases would be used in the analysis, it was decided that the coding of the scripts would have to be more open at the start to avoid omitting data. The second issue was that the data could be examined using quantitative measures once it had been categorised as this might well present new perspectives on what participants were doing while composing.

Another issue which had emerged during the pilot study was the pre-task used to prepare and train the participants. The e-mail writing task had taken too long for the native-speaker participants and was extending the amount of time that it took to carry out the report and interview to nearly two hours. A new pre-task was found using a jigsaw (see 6.3.3 for further information).

Other issues related to the frequency of the interventions and to the way in which the data had been transcribed. Interventions every five minutes meant that forty five minutes of writing only yielded nine opportunities for verbal reporting during the process. It also meant phases of writing were not being commented on because by the time the intervention came up the writer had forgotten what they had been doing at points during the five minutes. It was therefore decided to reduce time between interventions to

two minutes with the aim of providing a more detailed and comprehensive picture of how participants carried out the task.

5.11 Limitations of the product strand of the pilot and implications for the main study

Possibly the most serious limitation on the product strand of the pilot study was the very small number of scripts available for analysis. Such small numbers of scripts meant that statistical comparisons were very limited (see 5.5 above) and only descriptive statistics had been used for the pilot which meant that the findings could not be generalised to any wider populations. The initial process of getting candidates to produce both a B2 and a C1 level essay proved very demanding and time-consuming. It was therefore decided that for the main study scripts would be collected from the examinations themselves and that rather than the same titles being used at each level, the use of the same rhetorical pattern would provide parity. Section 6.5.4 in Chapter Six sets out the rationale for this in the main study.

Hyland's 2004 scheme had proved a useful tool for the analysis of the metadiscourse markers in the study but further reading showed that Hyland had further developed the list of metadiscourse markers as well as developing a more detailed definition of metadiscourse (2005). The extended list was useful for carrying out a more thorough search for metadiscourse markers but Hyland's definitions (p.37-48) also provided assistance in determining whether some instances of lexical exponents in the texts would qualify as metadiscourse or not, in particular the use of 'would' and other modals which the researcher felt had skewed some of the data.

5.12 Conclusion

The pilot study provided a valuable opportunity to trial some of the methods which would be used in the main study as well as identifying problems which would need to be overcome or managed. In that respect it

was a very useful exercise. However, the data obtained from the study was limited because of the small numbers of participants and the issues around the essay questions which had been used.

Chapter Six will set out the research methods for the main study and will also detail where appropriate how the limitations and implications of the pilot study were responded to.

Chapter Six: Research Methods for the Main Study

6.1 Introduction

This study follows the suggestion in Shaw and Weir (2007, p. 6) that for an argument of validation to be claimed, the process of writing and the products be explored. This has led to the creation of 'process' and 'product' strands to the study. The investigation into 'process' focusses on the verbal reports of participants carrying out the tasks while the 'product' strand focuses on the analysis of candidate scripts.

As has been stated in Chapter One, there are two aims of the investigation in this project. These are to answer the questions:

1. To what extent is **cognitive validity demonstrated** in the cognitive phases that candidates carry out while producing scripts at levels B2 and C1 in the English Speaking Board ESOL International Examinations?
2. What is the role of discourse competence in deciding whether a script is classified as being level B2 and C1 of the Common European Framework of Reference for Languages (CEFR) in candidate scripts from the ESB ESOL International Examinations?

The first aim is concerned with the cognitive phases which learners engage when writing timed essays. As discussed in Chapter Two, the cognitive phases cannot be observed directly. As a result verbal reports have been chosen to investigate how candidates produce their texts with the aim of gathering evidence that might be indicative of the different cognitive phases that candidates go through when writing their essays. This first aim is orientated towards the analysis of the process of writing.

The second aim focusses on the texts produced by candidates taking the examinations. This question examines how the distinction between B2 and C1 is evidenced in scripts and whether there are significant differences in the ways in which discourse competence is demonstrated between the two levels. As set out in Chapter Four, metadiscourse markers have been chosen in this study as the means to seek evidence of developing discourse competence in the collected samples of written work.

As identified in Chapter Five, it was decided in the pilot study that the two research questions were too broad in scope. The decision was then taken to retain both questions but to use them as 'aims'. As reported in 5.9.1, in order to investigate these two aims and to investigate 'process' and 'product' five research questions were developed to be addressed in the study.

6.1.1 Research Questions for the Study

Aim 1: investigation of processes:

1. What cognitive phases do candidates at levels B2 and C1 appear to employ when composing timed essay tasks?
2. To what extent does the timed essay question format used in the ESB ESOL International Examinations elicit the cognitive phases that models predict at levels B2 and C1?

Aim 2: Investigation of products:

3. Is there a difference in the quantity of metadiscourse markers used by candidates of the ESB International ESOL Examinations at levels B2 and C1 of the CEFR?
4. What are the functions of the metadiscourse markers used by candidates at level B2 and C1?
5. To what extent do the metadiscourse functions employed by candidates at levels B2 and C1 match the predictions made by the

CEFR regarding the development of discourse competence in learners at these levels?

By exploring process and product the study intends to explore how candidates draw on discourse competence when composing texts. This can then be considered through the texts themselves where the language used may indicate changes in the way writers carry out aspects of discourse competence in their writing.

Chapter Six will begin in 6.2 by considering the theoretical model within which the whole investigation is being conducted. The chapter will then outline the overall design of the process study (6.3) before setting out how the process data was analysed in 6.4. The research methods of the product study are described in 6.5 before the analysis methods for the product side of the study are described in 6.6.

6.2 Approach to the research

Section 5.1.1 of Chapter Five defined and made the case for the use of both quantitative and qualitative data in the study. In the current study, the use of the quantitative and qualitative methods is linked to the two aims of the study and the process and product strands. The first aim, looking at the process of writing involves the collections of qualitative data while the second aim, the focus on the products, will use a quantitative approach to look at metadiscourse markers as an indicator of discourse competence. However, the data from the verbal reports will also be partly analysed quantitatively in order to support the observations made and therefore demonstrate a degree of triangulation, or verification of the results (Dornyei, 2007, p. 164). The two strands will also be brought together to examine the extent to which cognitive phases implied by the data can be illustrated by changes in the way metadiscourse markers are deployed in scripts as evidence of developing discourse competence.

The next section (6.3) describes the design of the process strand of the study by discussing the participants, the tools used and issues related to the collection of the data.

6.3 Data collection: Process

The data collection tool in the process strand was qualitative in nature (i.e. the verbal reports). This section begins by describing the participants (6.3.1) before justifying the design of the verbal report format used (6.3.2) and then moves on to discuss the tools in sub-sections 6.3.3 to 6.3.7. the procedures used for the data collection are set out in 6.3.8 and the ethical considerations in 6.3.9. Note that the subjects in this strand of the study are referred to as participants as they were volunteers who took part in the study rather than examination candidates.

6.3.1 Participants

All twelve participants in the process strand of the study were Greek nationals with Greek as their first language. They were aged between 14 and 22. Greek nationals were chosen as the ESB ESOL International All Modes Examinations are used principally in Greece where around 20,000 candidates are assessed using the English language tests each year. The age range is also typical of the types of learners who take the assessments. As set out in Chapter One, the vast majority of these students study in 'frontistiria', the private language schools which learners attend in Greece.

Twelve participants took part in the verbal reports, six at B2 level and six at C1 level. All of the participants had successfully passed a test at the level they were intended to represent in this study including a pass in the writing section of that test and this was one of the key criteria for their eligibility. All of the participants had taken and passed either the Cambridge FCE (B2) or the Cambridge CAE (CAE). They were also in the age range of the majority of candidates who take the operational examination.

Table 6.1. *Overview of participants in the verbal report study.*

Level	Number of Participants	Nationality & First Language	Age range	Gender	Level substantiated by
B2	6	Greek	14-16	6 Female	Pass at Cambridge FCE including Pass in writing
C1	6	Greek	15-22	2 Male 4 Female	Pass at Cambridge CAE including pass in writing

The participants were from four different language schools across Athens and either signed the consent form (which was in Greek – see Appendix Six) or if they were under the age of eighteen had the form signed by their parent or guardian. The participants were provided two weeks in advance with an information sheet in Greek explaining the study and emphasising that the results would not be used for any purpose other than research and that the results would have no bearing on any future tests that they took.

6.3.2 *Design of the verbal reports*

The design of the verbal reports did not change a great deal from the tool for verbal reporting which was used in the pilot study (see Chapter Five, section 5.2.1). As in the pilot study, the main concerns were:

1. The participants should not be over-burdened by the requirement to report what they were doing;
2. The participants should have as much time for the task as they would under normal circumstances in the actual test;
3. As far as possible the implementation of the verbal report should avoid or at least reduce reactivity and veridicality.

In order to reduce the cognitive load on the participants the principle of a retrospective rather than a concurrent verbal report was retained from the pilot for the reasons given in Chapter Five (Section 5.2.3).

Other features from the pilot were retained such as the time allowed, candidates having no access to any resources (such as electronic dictionaries) that they would not have access to in the real examination and that the reviewer would not provide guidance on how to carry out the essay task (see 5.2.3 for further details). The same principles from Green (1998, p. 50) regarding the design of the tools were also followed.

The research tools for the verbal report in the final study comprised:

- a pre-task;
- an essay task;
- a researcher's script;
- a writing sheet for the candidate and question paper;
- a template for field notes to be used during the verbal report.

These tools are now discussed below in sections 6.3.3 to 6.3.7. The research tools each went through various stages of development, beginning with the pilot study for which the methods are reported in Chapter Five.

As explained in Chapter Five, the pilot had been run with native speakers carrying out the reporting tasks in order to trial different methods of reporting. As the verbal reporting tools had been adapted after the pilot (see the following sections) and had not been used with B2 or C1 learners, two further trial-runs were carried out. Two Chinese students from the University of Central Lancashire participated in the trials. One of these students had obtained a pass of 6.5 in IELTS and a sub-score of 6.0 for writing while the other had obtained a pass at 7.0 in the same examination (with a 7.0 in writing). This effectively indicated that the first represented a strong B2 level learner while the latter had achieved a clear C1 level (IELTS, 2012). The two volunteers went through the full process including the changed pre-task. Both participants were able to report on what they were doing and completed the writing task within the allotted time.

6.3.3 The Pre-Task

As set out in 5.2.4, pre-training had been conducted in the pilot study but the nature of task was changed for the main study. The training task used in the pilot study had consisted of asking the participants to produce a short e-mail and to retrospectively report on this at timed intervals, just as they would do in the main task. However, this activity had proved to be time-consuming, repetitive and complicated. It was also felt to be potentially intimidating for younger candidates so an alternative activity was sought.

Drawing on Johnson et al's (2008) study, the decision was taken to replace the e-mail writing task with a 12-piece jigsaw activity. This activity would reduce the amount of writing required by participants, speed up the pre-task training and would also be less threatening to the participants. It also ensured that the actual writing and reporting task could be completed in under an hour and the whole interview was concluded within an hour and fifteen minutes rather than the two hours that was required in the pilot study. Green (1998, p. 43) suggests that verbal reporting should not continue for more than an hour as the high level of cognitive processing

required will begin to impact on the participant and thereby cause problems of reactivity and veridicality as participants lose concentration. The two volunteers who the task were trialled on both responded positively to the task and it appeared to help them become more relaxed about the research they were involved in.

6.3.4 The essay writing task

In order for the study to be faithful to the examination which it was investigating it was important that the writing task was as close as possible to the types of tasks used in the exam. For this reason a previous exam title which had been approved by six raters at levels B2 and C1 was used for the verbal reports. The essay titles used for the process and product strands in the main study are described in 6.5.4.

6.3.5 The researcher's script

A version of the researcher's script had been developed for the pilot study. The script comprised the initial instructions to the candidate, the instructions for the pre-task, the main prompt to be used for the task, follow-up prompts to be used if required and the interview questions.

The development of the instructions went through a number of phases. As is reported in Chapter Five, the instructions were originally tested in the pilot study and were found to work relatively well in eliciting the verbal report. However, it was found that some participants had a tendency to simply report what they had been writing about, rather than talking about the writing process. A back-up prompt was added to the main prompt to be used if the researcher felt that the participant was overly focussing on the content rather than on the writing process. As a result, when such a participant was interrupted they were given the following instruction:

“Please stop and tell me what it is that you have been thinking about and doing in the last two minutes. You might want to talk

about the process of writing and any problems or difficulties you've been thinking about.”

The time between the prompts while writing was shortened from once every five minutes, as was used in the pilot study, to once every two minutes. This had the effect of more than doubling the opportunities for a candidate to report what they were doing and it was hoped that it would assist in providing a richer set of data than had been generated during the pilot study. The trial with the two Chinese volunteers demonstrated that the more frequent interventions resulted in more data and did not seem so frequent as to overly disrupt the writing process.

The instructions for the verbal reports for the main study were shown to a number of colleagues and there was also the opportunity to present the research tools to a number of researchers via post-graduate presentations at UCLan and also at CRELLA. An important contribution from the presentation at CRELLA was the instruction to the participants that they could choose to report in their native language if they wished. This step had been initially resisted by the researcher based on a findings reported in a study into essay writing by Plakans. Plakans (2009, p. 567) offered participants in her research the option of reporting their thoughts in their native language. She found that the take-up among the students was quite low but she also recorded that those participants who did choose to use their native language to report said that they found it very challenging to move between the two languages due to the process of translation. The decision was taken to inform participants that they could report in their L1 if they wished as the B2 level participants in particular might face difficulty in expressing what they had been thinking or to explain their thoughts in English. Also, the age of the participants was potentially quite young and therefore they might find the task too daunting. Both of these factors were felt to pose potentially serious threats to the study so all the participants were informed that they could express their

ideas in their L1 if they wished and the instruction was therefore added to the tool. In the event, neither of the volunteers nor any of the participants chose to report in L1 despite being offered the choice suggesting the Plakans was correct in her observations.

6.3.6 The writing paper and question sheet

As stated in 5.2.7, the layout of the paper on which the participants were to produce their essays was designed with two purposes. Firstly, it mirrored the appearance of the writing booklet that candidates use in the tests by being lined sheets with a margin. During the pilot and the trial of different reporting methods a larger than normal margin was provided on the left-hand side of the page for participants to record notes, ideas and comments. Some participants in the pilot had found this useful as it provided them with space to write down thoughts that they could talk about during the interruptions by the researcher. This feature was therefore retained for the main study.

Additional scrap paper for participants to produce plans or drafts was not provided since this would not be available in the actual examination. However, participants were told that they could write on the actual question paper (as many candidates do during the official examination) or that they could use the lined paper for this (again, many candidates do this in the real examination).

6.3.7 Field notes

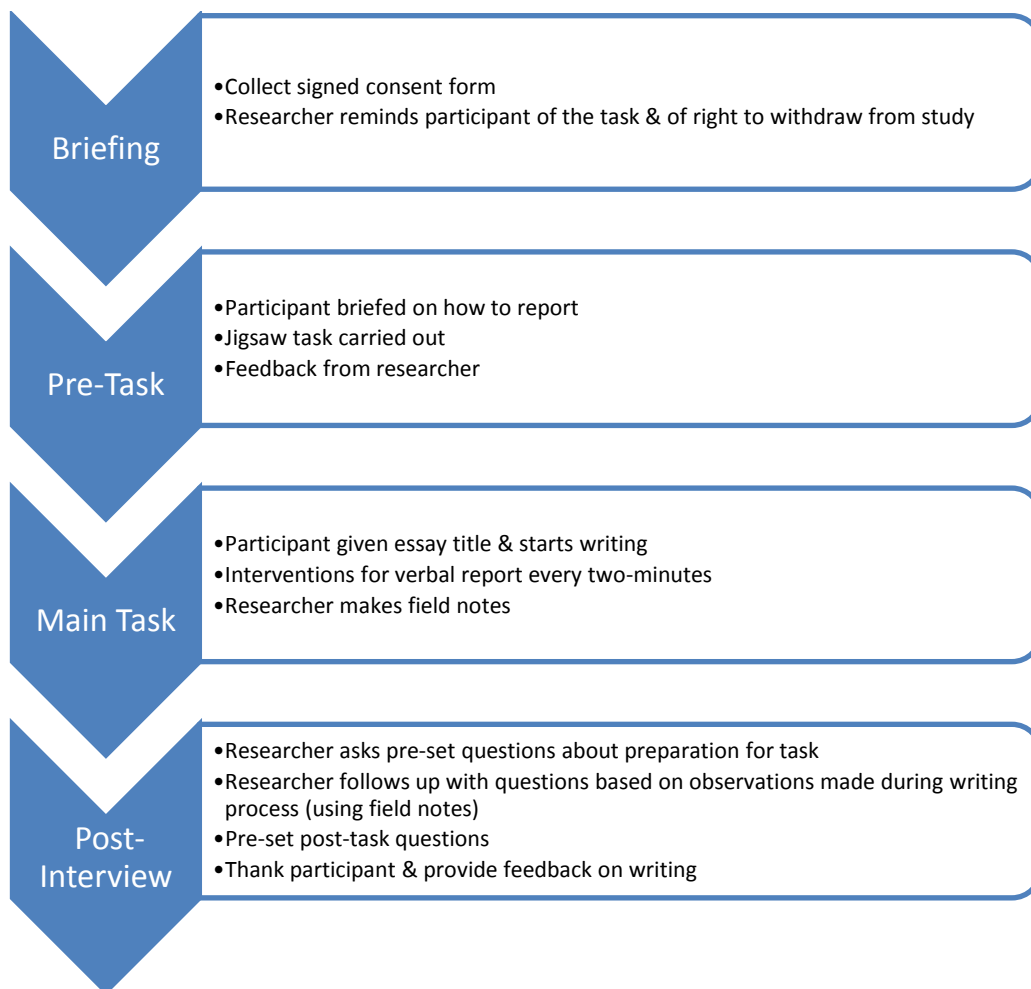
The use of field notes was a feature which very quickly emerged during the pilot study as the researcher found that it was necessary to record some of the details of what participants were doing during the timed-writing in order to ask about these activities in the post-interview. The use of field notes was therefore continued for the main study.

Field notes were kept using a running commentary with the time recorded of each observation on the left hand side of the sheet and the researcher's

observations and questions for probing in the interview on the right. The field notes were typed up afterwards (see Appendix Seven).

6.3.8 Procedures used for data collection

Figure 6.1 *Procedures for data collection in verbal reports*



The participants for the verbal reports were interviewed individually. The researcher began by explaining what would happen using the information sheet which had been provided to the participants two weeks in advance of the interviews and then the signed consent form was collected. The participant provided their most recent English language certificate as evidence of their language level.

The pre-task was carried out using the jigsaw puzzle task and any resulting questions regarding the verbal report process were answered. When the participant was ready to start the main task then the digital voice recorders were turned on. It was essential that the recordings of the verbal reports should be clearly audible as no attempt to transcribe on the spot was viable (Brown & Rodgers, 2002, p. 57). With this in mind, two digital voice recorders were used, one as a back-up in the event of a problem with the first.

The candidate carried out the writing task and then a short interview was conducted based on the pre-planned questions and the observations collected by the researcher using the field notes. The three pre-planned questions were:

- i. How did you prepare to write the essay?
- ii. What did you think about after you had completed the piece of writing?
- iii. When you are writing an essay like this in an exam, who do you think you are writing for?

Questions i and ii were asked at the beginning of the interview and were focussed on the writing task that the participant had just completed with the aim of exploring writing preparation and post-writing evaluation. Both questions had to fit with Ericsson and Simon's (1980, p.228) requirement that the prompts be based on the main task which the participants had carried out. Because question iii required the participants to consider their wider approach to writing it was the very last question asked in the interview and the responses were not considered to be part of the verbal report, nor were they used to inform the coding process.

Once the interview was concluded the data-recorders were switched off and candidates were offered some feedback on their writing.

6.3.9 Ethical considerations

Official channels were used to gain access to the participants (Cohen, Manion, & Morrison, 2000, p. 57). Information sheets were provided which included full information on the purpose of the research, the procedures to be used and anonymity. Permission to conduct the research was gained from the UCLan Ethics Committee. As part of this the researcher contacted the English Speaking Board and provided them with copies of the research tools and the information sheet. The researcher requested that ESB contact Europalso (the federation of language schools who are ESB's customer). A translated version of the information sheet and the consent form was provided for Europalso as it was possible those administering the request or the language school owners may not have had sufficient English to be able to understand the English version. Permission was obtained and the language schools identified participants and provided them with the information sheets and consent forms two weeks before the research was carried out. Where participants were below the age of eighteen, parents or guardians signed the forms (see Appendix Six).

The research was conducted within the language schools during their opening hours. The researcher went over the information sheet and emphasised to the participants that they could withdraw at any stage without any consequences or anyone being informed. When one participant expressed worries that his/her writing would be shown to one of their teachers they were reassured that that would not be the case and that the resulting data would be anonymised. The participant was also reminded that if he/she wished they could withdraw from the study without consequence. The names of participants were not used in the transcripts, only their level, gender, age and the researcher's own number for the participant were used thereby ensuring anonymity

The researcher felt that some kind of 'payoff' for the participants was essential because as Cohen, Manion and Morrison (2000, p. 57) put it

“people who agree to help are doing you a favour”. All of the participants took up the offer of feedback on their writing and a discussion with the researcher about exam writing strategies at the end of the interview. This was not recorded but all participants were keen to get the feedback and to discuss exam writing.

6.3.10 Transcription

As set out in 5.2.10, transcription is an important stage in the treatment of verbal reporting as the written report is used as the basis for the analysis. Green (p.51) identifies that in order for a transcript to be faithful it must retain paralinguistic features such as pauses and laughter and errors, mistakes and slips made during speech. Since the research in the main study focusses on language learners, some in the B2 level, it was reasonable to assume that there would be errors (incorrect use of language due to gaps in the learners knowledge of the language) mistakes (incorrect use of language which is ‘new’ or is still being accommodated within the learner’s interlanguage) and slips (performance errors). It is tempting, especially in the case of slips in particular, to ‘tidy up’ the transcript, since it is usually clear from the context what was meant. However, it is not always possible to identify which type of error has been made meaning that the resulting transcript would be in danger of becoming a subjective interpretation of the verbal report, rather than an accurate and faithful transcription. Therefore, errors were included in the resulting transcript.

The transcriptions were written up using the transcription convention from CANCODE (Adolphs, 2008, p. 137-138). The CANCODE format was chosen because it provides a high level of readability for the resulting transcript and it is also compatible with a wide range of software applications such as Wordsmith and NVivo (QSR International, 2014). In the pilot stage, the transcription had omitted many paralinguistic features and it had been felt that it had not been segmented sufficiently to allow for

effective analysis (see 5.10). However, as the data was collected primarily for the purpose of identifying possible cognitive phases drawn on by candidates during the writing, few meta-linguistic features needed to be identified (i.e. intonation, stress etc).

The transcriptions were carried out by a four individuals. The researcher ensured that all of those involved were provided with a sample transcript and the conventions (based on one of the two pilots of the main data collection tool). The researcher also checked the completed transcriptions against the recording for accuracy and to ensure that the transcribers had been consistent in their use of the transcription code.

6.4 Data analysis: Process

Having reviewed the data collection method for the process strand of the study, this section now goes on to discuss the analysis methods for the data from the process strand of the study before moving on to discuss the data collection for the product strand of the study in 6.5

As discussed in 5.10, the use of verbal protocols had been chosen to investigate the ways candidates at levels B2 and C1 of the CEFR compose their texts. The aims of the process strand of the study were to see which cognitive phases the essay writing tasks elicited and whether the cognitive phases used by the candidates were those predicted by the CEFR. In order to do this, two research questions were posed for the process strand of the study:

1. What cognitive phases do candidates at levels B2 and C1 appear to employ when composing timed essay tasks?
2. To what extent does the timed essay question format used in the ESB ESOL International Examinations elicit the cognitive phases that models predict at levels B2 and C1?

As set out in 2.6, Field's (2004) model was chosen for comparison because of its use by Shaw and Weir (2007) to examine the Cambridge ESOL suite of examinations. In addition to Field's description of the cognitive phases, Shaw and Weir describe the various processes (p.44-62) which each level of assessment is expected to produce as well as observations gained by examiners on features of the texts. This provides a model against which the current study can consider its results.

6.4.1 Coding

Brown and Rodgers (2002, p. 63) identify that the process of coding is a critical one when looking at qualitative data collected from interviews or verbal reports. Qualitative data is likely to be "messy" (Dornyei, 2007, p. 244) and in order for it to be useful it has to be given shape and organisation. Dornyei (2007, p. 250) distinguishes between two processes, 'pre-coding' and 'coding'. Pre-coding comprises the reflections of the researcher as they read and re-read the data. This then gives way to a more formal coding process in which categories are made more concrete through definition. Green (1998, p. 73) argues that whatever coding scheme is produced it should reflect the following points:

- The coding scheme must be able to capture as much data as possible from the verbal reports;
- The coding scheme should be relatively free of theoretical assumptions to allow data to emerge;
- The coding scheme should enable the researcher to test hypotheses which are consistent and inconsistent with the approach;
- The coding scheme should allow for the variations which will occur between participants as they do the tasks.

While the pilot study made use of Fields (2004) cognitive phases as an *a priori* scheme to code the data, this was felt to have been problematic (see

5.10) as it may well have excluded data (see Green 1998, p. 70). For the main study the data was not coded against a pre-existing scheme but was later considered against Field's cognitive phases (Field, 2004). It was felt that it was important to consider the data from the verbal protocols in its entirety in the pre-coding stage so as to allow categories to emerge. However, as Dornyei points out (2007, p.253-254), even at the pre-coding stage, the researcher's bias can influence the way that coding is approached, so Field's phases may still have influenced the analysis.

The data from the verbal reports took three forms: the verbal report from the task, the immediate post-interview and the field notes. It was decided that the verbal report would be used as the main data source with the post-interview and field notes used to interpret or clarify what a participant had reported. As discussed in 2.5.2.2, retrospective interviews are likely to be accurate and free of veridicality if they are conducted directly after the verbal report, are based on the actual report that participants have carried out (as opposed to talking about other instances) and if the investigation focusses on heeded information rather than requiring justifications. The interviews featured a set of questions which asked the participant about issues which went beyond the scope of the immediate task (i.e. whether they had participated in similar research before and who they usually addressed essays to when they were writing) but these questions were asked right at the end of the interview and were not used to assist in the coding process.

Segmentation is the process of dividing the text from a verbal report into the units according to which it will be analysed. Texts are usually divided for analysis into phrases, clauses or sentences (Green, 1998, p. 73) with the aim of making coding easier. For this study, it was found that most of the responses in the verbal report were fairly short, comprising one or two sentence-length utterances. Therefore, each intervention by the researcher (every two minutes) was regarded as a segment due to the

participants' responses being quite short. This also allowed for some comparison between participants regarding when different activities were mentioned by the participants because the interruptions were standardised.

6.4.2 Stage 1: Initial coding

Following Green's recommendation that the coding should begin from as theoretically free perspective as possible (1998, p. 73) the first stage was a read-through of the verbal protocol data with any noteworthy passage highlighted and a coding tag written in the margin beside it. This was done so as to allow previous unconsidered elements to emerge (Dornyei, 2007, p. 251) but it also necessitated that the coding be as clear as possible both to allow for the identification of other instances of the same feature in the scripts but also to identify when a new code was needed. Table 6.2 sets out the categories which emerged from this initial coding. The process also helped to meet Green's other stipulations in that the pre-coding process was carried out with the intention to capture as much data as possible whether it was 'relevant' to the research questions or not. The pre-coding exercises also meant that the categorisation process was constantly tested as features were identified.

As can be seen from Table 6.2, allowing codes to emerge from the data allowed for a much wider consideration of the data than would have resulted from the application of an *a priori* scheme as was used in the pilot.

The initial coding process identified and then refined categories. For example, a category was initially created which was termed 'Local Planning' based on comments by participants about planning the section of text immediately about to be produced. This category also included instances of participants reporting that they were thinking about how to link what they were writing to the next paragraph. As coding went on, it was felt that this was a different level of complexity from simply ordering the

content of a single paragraph because it implied a notion of textual cohesion beyond the sentence level. This observation resulted in the creation of a new category named 'linking paragraphs'. A second example of an emerging category was 'Monitoring'. The two categories 'Monitoring Content' and 'Monitoring Language' emerged quite early in the pre-coding process but it was later found that there were instances where the learner talked about re-reading their text without specifying why and the post-interview questions failed to ask what the participant had been doing. Rather than second-guessing the purpose of the re-reading, the category of 'Monitoring' was created which covered the process of a participant re-reading without mention of the specific purpose.

Table 6.2. *Categories identified from the initial coding.*

Category Title	Definition (based on initial exploratory coding)
Task Assessment	Consideration of the set task and its parameters. Explicit reference to the set task. This includes attempts at summarising the essay question.
Generating Content	Participant is coming up with ideas for content to go into the essay based on their own knowledge and experience.
Text Level Organisation	Participant is considering how ideas are to be set out across the whole essay.
Local Organisation	Participant is considering how to organise the ideas in a particular paragraph.
Linking Paragraphs	The participant explicitly considers how to connect two paragraphs in the text.
Immediate Planning	The participant is considering the section of text which is immediately to be produced or which is in the middle of production.

Category Title	Definition (based on initial exploratory coding)
Writing	Participant reports what they have just written but without discussing how it is shaping what will come next.
Stuck	The participant reports being 'stuck' but does not specify the source of the problem.
Searching for Lexis	The participant reports that he/she is mentally searching for a particular word or phrase.
Monitor	The participant reports re-reading but without specifying the purpose of the re-read.
Monitoring Content	The participant re-reads with expressed intent of evaluating whether the ideas used are appropriate, relevant or accurate.
Monitoring Language	The participant re-reads in order to identify errors, mistakes or slips in the language they have used.
Revision	The participant makes changes to the content or the language based on re-reading.
Summarising Content	The participant is attempting to find a way to summarise what they have already written.
Consider Audience	The participant talks about the impact that their writing will have on the reader or for whom they consider the essay is being produced.
Word Count	The participant counts or considers the total length of the text.

Once the initial coding was completed a colleague was asked to use the code on a clean copy of two of the scripts. The raw results of this are given below in Table 6.3 while the results of agreement by category are given in Table 6.4. An agreement rate of at least 80% was sought in line with Green's recommendations (1998, p. 19). Where the two raters' coding disagreed (as in the case of Immediate Planning, Local Organisation and Writing) discussion took place to compare results and see whether the issue could be addressed.

The main sources disagreements between the researcher and the rater were over the categories of Immediate Planning, Local Organisation and Writing. What became clear in the post-discussion was that the boundaries of these categories were unclear. Table 6.3 illustrates that for

script A in interventions 5 and 7 the researcher and rater applied different categories. This can also be seen in script B in interventions 5 and 11.

Table 6.3. *Researcher and rater trial of the coding categories.*

Interventions (2 mins)	Researcher script A	Rater Script A	Researcher Script B	Rater script B	Key
Int 1	TA; O	TA; O	TA; IP; GC	TA; IP; GC	GP = Generating Content
Int 2	LP	LP	GC	GC	IP = Immediate Planning
Int 3	GC	GC	GC	GC	LO = Local Organisation
Int 4	GC	GC	GC	GC	LP = Linking Paragraphs
Int 5	IP	LO	W; GC; IP	W; GC; LO	MC = Monitoring Content
Int 6	S; LO; GC	S; LO; (GC)	GC	GC	ML= Monitoring Language
Int 7	W; O	IP (O)	GC	GC	N/A = No Applicable Comments
Int 8	GC (TA)	GC; TA	GC	GC	O = Organisation
Int 9	S	S	W; GC	(W) GC	R = Revision
Int 10	GC; LO; IP	GC; LO; IP	GC	GC	S = Stuck
Int 11	MC	MC	IP	LO	SC = Summarising Content
Int 12	ML; R	ML; R			SL = Searching for Lexis
Int 13	SL; S	SL; S			TA = Task Assessment
Int 14	IP	IP			W = Writing
Int 15	N/A	N/A			
Int 16	MC; R	(MC; R)			
Int 17	LO; SC	LO; (SC)			
Int 18	SL; SC; TA	SL; (SC)			
Int 19	SC; SL	LO			

Codes in parenthesis were agreed in post-coding between raters

Table 6.4. *Agreement by category.*

Category	Script A			Script B			Mean difference
	# of instances	# of differences	Agreement (%)	# of instances	# of differences	Agreement (%)	
Generating content	10	1	90	20	0	100	3.33
Task Assessment	5	1	80	2	0	0.00	14.28
Organisation	4	0	100	0	0	100	0.00
Immediate Planning	6	2	77	4	2	50	40.00
Local Organisation	8	2	75	4	2	50	33.33
Writing	1	1	0.00	4	4	0.00	100.00
Monitoring Content	4	0	100	0	0	100	0.00
Summarising Content	5	1	80	0	0	100	20.00
Searching for Lexis	5	1	80	0	0	100	20.00
Reviewing	4	0	100	0	0	100	0.00
Stuck	2	0	100	0	0	100	0.00

It emerged from the discussion that the underlying issue was the difficulty in defining the difference between Immediate Planning and Local Organisation since both categories concerned the imminent production of text and implied some level of local organisation. Field (2004, p. 329) notes that planning at a paragraph or sentence level will involve consideration of the local level organisation but also will involve syntactic

concerns such as how to sequence the ideas in the piece of text to be produced. An example of this can be seen in the transcript of participant 5 (B2) who was asked in the post-interview about why she had crossed out some of what she had written.

“I here had write one positive impact reason but and I want to write something else people er lived that lived years ago and they <\$=> I wasn’t <\$E> inaudible <\\$E> with the other and so I cross it. I want to begin that with yes people er years ago er they didn’t know the live as much as they do now. So it wasn’t <\$=> wasn’t make sense to write it that way.” (Participant 5, B2 <36:06>).

The participant suggests that a decision to include additional content, made at the time of writing (Generating Content) then became a matter of local organisation but required syntactic changes to the language being produced (Immediate Production).

Shaw and Weir (2007, p. 53-59) review the cognitive phases used by candidates in the Cambridge ESOL suite of examinations and interestingly enough choose to include micro-planning as part of the wider organisation category. They also comment that translation, that is transformation of mental ideas into consensual codified text, is “not susceptible to direct investigation” (p. 57) because the process is largely automatised. The comments from participant 5 above also emphasise the iterative nature of writing and show how actual production also includes constant monitoring and revision phases.

As a result of the difficulties that the researcher and the rater experienced in differentiating between Local Organisation and Immediate Planning, both in their individual rating and in agreeing in the post-discussion, the decision was made to collapse both categories into the category of Microplanning, which would reflect Field’s definition (2004, p. 329) and would bring rater agreement to the 80% suggested by Green (1998). See Table 6.5 below.

Table 6.5 *Cognitive phases in the final rating scheme.*

Final Categories for coding	Categories from initial coding scheme
Task Assessment	Task Assessment
Word Count	Word Count
Generating Content	Generating Content
Organisation	Text Level Organisation
Microplanning	Local Organisation
	Linking Paragraphs
	Immediate Planning
	Writing
Stuck	Stuck
Searching for Lexis	Searching for Lexis
Monitor	Monitor
Monitor Content	Monitoring Content
Monitoring Language	Monitoring Language
Revision	Revision
Summarising Content	Summarising Content
Consider Audience	Consider Audience
Time Pressure	Time Pressure

An attempt was made to preserve the Writing category and it was redefined so that it only referred to explicit instances of the participant reporting what they had already produced. This meant that Microplanning as a category focussed on looking ahead whilst the Writing category was invoked when a participant said “I wrote/have written...”. However, once coding was started it proved difficult to clearly identify instances. For example, when ‘I have written’ or ‘I am/was writing’ are used by participants, it was hard to decide whether what they had produced was complete or ongoing so the division with the Immediate Planning category

became increasingly fuzzy. Finally, the Writing category was scrapped with all such instances being put into Microplanning as they concerned what was to be produced but also concerned how what had already been produced was influencing the shape of the text. The final rating scheme is given in Table 6.5.

6.4.3 Coding into NVivo

With the categories agreed the scripts were then recoded by the researcher using NVivo software (QSR International, 2014). This was done without reference to the annotated scripts so as to re-test the original coding exercise and to demonstrate intra-rater reliability on the part of the researcher. While Green (1998, p. 93) suggests that this is a process which has faults, the coding system had already been tested with a second rater in stage one and this was an opportunity to see whether the researcher was applying the coding consistently himself. The level of intra-rater consistency was found to be very high with few post-adjustments. The coding exercise was carried out three weeks after the initial coding and the adjustments to the categories in order to reduce the chance of the researcher simply remembering how the text had been coded previously.

6.4.4 Field's cognitive phases

With the verbal report scripts coded, the categories from stages one and two were then matched against Field's (2004, p. 329-331) cognitive phases using both Field's descriptions and the interpretations in Shaw and Weir (2007, p. 34-62) to create a 'tree' of category codes (Dornyei, 2007, p. 252-253). This meant that Field's phases were used as superordinates and the categories identified in the script analysis were used to further described Field's phases. Categories which did not fall into Field's phases remained outside of the scheme for separate analysis. Subdividing Field's phases was a level of analysis which had been identified as lacking in the pilot study (see 5.10). Table 6.6 shows how the categories developed from the initial coding were mapped onto Field's cognitive phases using

the descriptions in Field (2004) and Shaw and Weir (2007). This division was also based on the discussions with the 2nd rater (See 6.4.2 above) in that where categories could not be adequately separated, they were included under one of Field's broader phases, as is the case in Microplanning.

Table 6.6 Mapping of categories onto Field's cognitive phases

Cognitive Phase	Categories	Examples from scripts
Macroplanning "the writer assembles a set of ideas, drawing upon world knowledge...establishes what the goal of the piece of writing is to be...target readership...genre...and style." (Field, 2004, p.329)	Text Assessment (Field, 2004, p.329; Shaw and Weir, 2007, p.38)	"I read very well er all the task er I try to understand exactly what it wants" (Participant 1, C1)
	Generating Content (Field, 2004, p.329; Shaw & Weir, 2007, p.38)	I've been thinking of the ideas. I've been brainstorming on what I can write (Participant 3, C1)
	Word Count Included as aspect of macroplanning as it concerns expectations of readership.	I was afraid that it er will be too much and I will er pass the...word limit (Participant 2, C1)
	Consider Audience Consideration of readership	I'm just trying to think [ideas]...that er...once someone reads it er it makes a difference in all the other er essays (Participant 1, C1)
Organisation "the writer organises the ideas, still in abstract form" (a) in relation to the text as a whole and (b) in relation to each other...outcome may be a rough set of notes." (Field, 2004, p.329)	Text Level Organisation Organising ideas although still on the abstract level for the whole text but also in relation to each other (Field, 2004, p.329)	I wanna say how will I separate the paragraphs...I think it will take three paragraphs for this information...I'm trying to put in order all my thoughts....and all the reasons..." (Participant 1, C1)

<p>Microplanning “the writer shifts to a different level and begins to plan conceptually at sentence and paragraph level. There is constant reference back to...decisions made at earlier stages and to the way that the text has progressed so far...the writer needs to give consideration to whether an individual piece of information is or is not shared with the reader.” (Field, 2004, p.329)</p>	<p>Local Organisation Microplanning can concern the goal of the paragraph considered against what has been produced and what is about to be produced and what information has been given, what is already shared and what might need adding to clarify what is meant (Shaw & Weir, 2007, p.39, p.55)</p> <p>Consideration of the “text so far” (Shaw & Weir, 2007, p.39)</p> <p>Pre-consideration of sections which are not about to be immediately produced. (Shaw & Weir, 2007, p.53)</p> <p>Immediate Planning Pre-planning at sentence level, often with reference to paragraph level and macro-goals (Field, 2004, p.329)</p> <p>Linking Paragraphs Consideration of text direction and the fit of the current section against the organisational plan (Field, 2004, p.329).</p> <p>Summarising Content Consideration of what information has been given, what is already shared and what might need adding to clarify what is meant (Shaw & Weir, 2007, p.39, p.55)</p>	<p>I’ve been developing the second er <\$=> my second paragraph and the first question. (Participant 3, C1)</p> <p>how I will start erm to <\$E> pause <\\$E> prepare who is going to read the essay about what I am writing in this paragraph (Participant 11, C1)</p> <p>What came in my mind is <\$E> pause 4.5 seconds <\\$E> the conclusion <\$=> I don’t know why but how to end the essay <\$=> it’s too early but I don’t know just pop in my mind.(Participant 6, C1)</p> <p>I’m still thinking of how to write the first paragraph in order for me to be more easily=easier to continue to the second paragraph (Participant 6, C1)</p> <p>I’m trying summarising everything in the conclusion (Participant 4, B2)</p>
---	---	--

<p>Translating Turning internalised text into public text (Field, 2005) and using consensual symbols (Kellogg, 1994, p.6) . Shaw & Weir (2007, p.40) identify that this is the point at which a learner may realised that they do not have the necessary linguistic resources. Micro-level decisions are made concrete at this point.</p> <p>Using key words to use in the essay from the planning stage (Field, 2004, p.329)</p>	<p>Searching for lexis Searching for the right piece of lexis would fall into the translating category (Shaw & Weir, 2007, pp.57-59)</p>	<p>I'm actually trying to enrich it with er good vocabulary (Participant 3, C1)</p> <p>I'm trying to find some good words and not just thinking of boring words to put in er my es= essay) Participant 1, C1)</p>
<p>Monitoring "is a complex operation at many different levels...checking accuracy of spelling, punctuation and syntax...examining the current sentence to see how clearly it reflects the writer's intentions" (Field, 2004, p.330)</p>	<p>Monitoring Re-reading of text (without specified reasons given)</p> <p>Monitoring Content "At a higher level, monitoring should involve consideration of the extent to which the text produced accords with the writer's goals, its relevance and adequacy for the set task and the development of the discourse structure of the text" (Shaw & Weir, 2007, p.41)</p> <p>Monitoring Language "Checking accuracy of spelling punctuation and syntax" (Field, 2004, p.330)</p>	<p>now I'm going to <\$E> pause <\\$E> read again (Participant 9, B2)</p> <p>I'm...elaborate too much this point and maybe I should short it a bit. (Participant 2, C1)</p> <p>I checked my punctuation throughout the paragraph (Participant 2, C1)</p>
<p>Revising "After monitoring...a writer will return to aspects of the text which he/she feels to be unsatisfactory and revise them...many of the revisions are at the lexical level" (Field, 2004, p.330).</p>	<p>Revising The decision not to correct if monitoring has identified to a candidate that something wrong is a revision decision (albeit a negative one)</p>	<p>I wanted to check if everything is okay for example some stupid mistakes I do when I write and I fixed them. (Participant 1, C1)</p> <p>My introduction word...it is ... not correct...but I left it like that (Participant 2, C1)</p>

The data from the post-interviews was used to assist in the coding and then used to interpret the findings from the verbal reports. NVivo (QSR International, 2009) was used to analyse the data and generate the results. The software was used because of its ability to allow coding of categories and facilitate some quantitative exploration of the data including word frequency (see 7.3).

6.5 Data collection: Product

This section now turns to the data collection methods for the product strand of the study. The data collected for this strand is more quantitative in nature in that it aims to assemble a mini corpus of candidate essays at the levels under investigation and then analyse the use of metadiscourse markers through a range of quantitative measures. 6.5.1 discusses the rationale for this strand of the study. Sections 6.5.2 to 6.5.8 go on to explore the design of the product study including how candidate scripts were selected for inclusion in the corpus. 6.5.9 discusses the ethical issues related to the data collection for this strand.

The scripts here are referred to as ‘candidate scripts’ and the writers as ‘candidates’ to distinguish them from the participants in the process strand of the study.

6.5.1 The case for quantitative data

Just as discussed in the pilot study (5.4.1), metadiscourse markers were chosen as the focus for the investigation into discourse competence for the following reasons:

- metadiscourse has interactive and interactional functions. The interactive functions include decisions regarding text organisation on the part of the writer. Interactional functions are those by which the writer seeks to intrude into the text and engage with the reader (Hyland K. , 2005, p. 218-224). Metadiscourse therefore may offer insights into how a writer is consciously shaping a text and taking

into account the expectations of the reader. In doing this the product strand of the study offers another perspective on the composition processes investigated in the process strand;

- metadiscourse provides a link from the macro-levels of text development (i.e. how a text is to be organised to best fulfil its intended purpose) to the micro-level of the linguistic exponents used. This suggests that metadiscourse may indicate how candidates manage the interaction with the **reader** (in terms of the interactive and interactional functions) through their choice of linguistic exponents.

6.5.2 *The design of the study*

The design of the study followed the principles set out in 5.4.1. The product strand of the study is an *a posteriori* study of candidate performance using a small-sized corpus. The limitations of the pilot study meant that more care was taken to control and verify the texts which were entered into the corpus. This selection and verification process is detailed below (see 6.5.4 and 6.5.6). As in the literature review, Cobb's (2003) questions are used in the following discussion to set out how the corpus was assembled.

Cobb (2003, p. 394)

- i. How large does this corpus need to be?
- ii. What types of text have been entered into the corpus?
- iii. By whom have these texts been produced?
- iv. How have these texts been 'approved'?
- v. How have the texts been transcribed and coded?

6.5.3 *Corpus size*

As discussed in 5.4.2, the size of a corpus depends on the purpose to which it is to be put and there are no clear rules regarding how large a corpus actually needs to be Baker (2010, p. 95). Of more importance, at

least according to Baker (2010, p. 96) is the variety of text types that make up the corpora as these determine the output results.

One of the serious limitations of the pilot study was the small number of scripts entered at each level. While small-scale studies are common in research into metadiscourse markers (Intaraprawat & Steffensen, 1995, p. 95), they are unlikely to be representative of the test-taking population and statistically significant results would need to be reported with caution. It was therefore intended that the main study would assemble a larger corpus of texts at each level which would allow for more solid conclusions that the pilot study was able to achieve.

Given the scope of the study, the size of the corpus that could be assembled was inevitably limited. The experience of piloting suggested that it was also important to exert greater control over the content. In terms of the main study, it was determined that the corpus should be of one type: the timed-academic essay and be restricted to one rhetorical question format: the advantage and disadvantage essay. All the essays should come from one defined population: Greek candidates of ESB ESOL International Examinations.

6.5.4 *Text types in the corpus*

The choice of the advantage and disadvantage essay was informed by the CEFR. At the B2 level that a learner should be able to “explain a viewpoint on a topical issue giving the advantages and disadvantages of various options; construct a chain of reasoned argument” (Council of Europe, 2001, p. 35). Appendix C of the CEFR, the Dialang scale for writing, also identifies that at C1 the candidate should have the ability to “develop an argument systematically, giving appropriate emphasis to significant points, and presenting relevant supporting detail” (p.232). Argumentation then is a shared mode at B2 and C1, so it was decided that the essays for the

corpus should have this rhetorical pattern in common in order to provide some parity.

The choice of the essay as a task was determined by the fact that advantage/disadvantage argumentation patterns are a common feature of the genre of essay writing. It was also felt that many of the test-takers use the examinations to develop their academic writing skills in English and as such the essay is a genre which they expect to produce and is also in keeping with the style of writing that might be expected in academic settings.

However, as detailed in 5.11, issues with the number of candidate scripts available at the C1 level in particular meant that it would not be possible to gather enough scripts on the same topic to satisfy the requirement for a larger corpus. Instead, the decision was taken to include essays with the same rhetorical pattern (advantage /disadvantages) but on a range of topics.

As set out in the research questions in 6.1.1, the purpose of this study for which the corpus was assembled was to identify which metadiscourse markers were used by candidates in the ESB ESOL Examinations at levels B2 and C1, explore the functions used and examine whether the CEFR's predictions were correct. As set out in Chapter Four, metadiscourse markers as defined by Hyland (2005, p. 38) do not carry propositional information and are concerned with interpersonal functions: interactive (i.e. assisting the reader in navigating their way through the text) and interactional ("involving the reader in the text") (ibid). Therefore it can be argued that the actual content of the essays was largely irrelevant so long as the rhetorical pattern of the essay questions remained the same. This lack of parity in the rhetorical pattern was an issue identified in the pilot (see 5.9.2) and without the uniformity of the pattern of the questions (i.e. if some tasks were 'compare and contrast' or

‘problem/solution’), it would be hard to determine whether responses were being influenced by the rhetorical demands of the question or by the level of the candidate. However, the decision to use different questions, so long as they fit the criterion of being (a) suitable for the level and (b) phrased with the same rhetorical patterns, could be defended.

It can also be argued that having a range of essay questions in the study may be a useful feature which gives the study more applicability. Were only one title to be used, it could be argued that the findings apply only to that particular essay question and that the study would be more akin to the development of a primary trait scale for the task (Fulcher, 2010, p. 208) rather than being generalisable to other tasks with the same rhetorical format.

In setting and selecting the questions for B2 and C1 a number of principles were followed in order to ensure that the tasks were appropriate for the levels at which candidates were being examined. As has been set out above, the CEFR identifies argumentation and being able to identify the pros and cons of a topic as being something that learners at the B2 level should be able to do and the framework suggests that learners at the C1 level should be able to extend this into “well structured expositions of complex subjects, underlining the relevant salient issues” (Council of Europe, 2001, p. 62).

The question used to collect candidate samples was one of a number of questions circulated to a panel of raters, all of whom were experienced users of the CEFR. The panel were asked to rate each proposed question in terms of the level they thought it was best suited and also to comment on whether they thought the question was appropriate or not. The task which was selected from this exercise was the only one to be unanimously agreed by all the raters to be both C1 in level and suitably phrased. Comments tended to agree with those of one of the raters who stated that

“the first part (the issue of ‘why’ something has occurred) gives the writer something to latch onto in order to build the positive and negative impacts.”

As has been argued in 3.7, since discourse is highlighted by the CEFR as being a feature of production which learners at the B2+ and C1 levels attend to in more detail, it is logical to conclude that candidates writing in the tests will show more awareness of the need to write for a particular audience. However, one feature of the essay questions in the examinations under review was that they do not overtly specify an audience and it is left to the candidate to interpret for whom they are writing. This feature of the task was left intact since the project is considering the role of discourse competence through the ESB examinations. However, a question was added to the process strand of the study (the verbal protocols) in which candidates were explicitly asked about for whom they were writing their essay in order to explore the issue.

6.5.5 Source of the texts

The Greeks make up the largest national candidature for the ESB ESOL International Examinations and could provide large numbers of essays for the study. The samples were collected from timed English language examinations at levels B2 and C1. Candidates in Greece took the examinations under the same conditions, as set out in the extensive examination handbooks and standardisation guides provided by UCLan (2014) to ensure that the test was administered under uniform conditions in all centres to support reliability (Brown, 2004, p. 34).

6.5.6 Approval of the texts

All of the scripts entered into the corpus went through a number of tests in order to identify that they were at the appropriate level for the population they were intended to represent (i.e. B2 or C1):

- All of the scripts were taken from candidates who had achieved a pass overall at the level.
- All of the writing scripts had been marked and awarded a pass grade in the writing section of the ESB ESOL International Examinations by standardised raters as part of the normal examination process;

Raters who are marking the examinations go through a number of quality measures in order to verify that their grading remains consistent in terms of inter-rater and intra-rater reliability. The rater-standardisation consists of the following:

- i. A pre-task aimed at familiarisation with the CEFR and the statements on writing.
- ii. Pre-marking access to criteria documents and the 'ground rules' for the marking.
- iii. A standardisation session for each level at which raters will assess. The standardisation involves marking a set of pre-rated papers. The grades assigned by the rater are then checked against the official grades and any issues/discrepancies are discussed.

In addition to the standardisation, raters are also moderated during the marking process in the following ways:

- i. Spot checking and second marking of a minimum of 20% of the rater's work is carried out across the marking period. Consistency is recorded and tracked. Where a rater is found to be inconsistent in their marking, they are re-standardised. Should a rater to be found to be unable to mark consistently despite interventions, they are removed from the marking.

- ii. Where there is a discrepancy of more than one band between the first and second marker, a third rater will be brought in to determine the final grade. This procedure is automatically carried out where there is a discrepancy in terms of a pass or fail grade.

All of the writing scripts selected for the corpora were marked according to the processes set out above. However, because the research questions not only focus on the functions carried out by candidates in the ESB ESOL International Examinations but also seeks to examine the CEFR's predictions about how learners at B2 and C1 use metadiscourse extra steps were added. In order to verify that the scripts were appropriate for the CEFR level which they were intended to represent, an additional rating procedure was carried out which is detailed in section 6.5.8 below.

6.5.7 Transcription and coding

The same procedures for transcription and coding were followed as were used in the pilot study (see 5.4.6). Texts were re-read by the researcher upon transcription and compared to the original hand-written text to ensure that Word's autocorrect feature had not changed any words and to ensure no text had been missed. Any inconsistencies were corrected at this stage.

6.5.8 Independent verification of script levels

The research questions for this study seek to explore the role of discourse competence as an indication of a candidate being at level B2 and C1 in the ESB ESOL International Examinations. The research also aimed to examine how well the CEFR predicts the development of discourse competence at these levels through the analysis of metadiscourse markers. It was therefore important that the candidate scripts selected should not only represent passing performances at the required levels of the English language tests under examination but also that they should be

representative of the relevant CEFR levels. To confirm their CEFR level, the scripts were rated by a second panel of raters with experience of using the CEFR. These raters were all ESOL-qualified and had experience as examiners for IELTS, Cambridge ESOL and other examining organisations.

In the first stage, the raters were required to complete an online CEFR benchmarking task (SurveyMonkey, 2014) in which they were asked to rank statements from the CEFR. These statements were taken from a modified version of the written assessment grid presented on page 187 of the Council of Europe's manual for *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching and Assessment (CEFR)* (2009). The same statements were used to develop the scheme which would be used by the raters. The scheme was adapted for the rating of essays and the levels B1+ and B2+ were added (see Appendix Eight).

Rasch analysis was used in order to ensure rater consistency in the independent verification of scripts. The advantages of using Rasch were that the model does not require all the scripts to be rated by all the raters so long as there is overlap between the scripts raters are assigned (Shaw & Weir, 2007, p. 192; Bond & Fox, 2007, p. 148). Rasch is able to assess consistency in the performance of raters as well as severity or lenience (Bond & Fox, 2007, p. 157). A rating matrix was drawn up to ensure that there was sufficient overlap in the scripts raters marked to allow the programme to measure rater performance.

Each rater was given 23-25 scripts to rate using the criteria. The data was then analysed using Minifac, a free version of FACETS software (Linacre, 2008). The results of the independent rating for the raters are set out in **Table 6.7** below. The results showed that all of the raters were marking consistently as their scores for the Infit Mean Square were between .6 and

1.4 as Linacre (1989) suggests for rating scales. Scores above this would have indicated a lack of consistency on the part of raters. Rater 2's grades were more severe than others, although the rating of candidates' ability was consistent. What was also apparent was that some of the scripts rated by Rater 2 had also been marked by raters 11, 7 and 1 all of whom had been slightly stricter than other raters. Use of FACETS was able to identify this and assisted in dealing with the problem of a low return in numbers of C1 scripts.

Table 6.7 *Rater Measurement Report (Arranged by mN)*

Rater ID	Average Rating	Fair-M Average	SE	Infit MSE	Outfit MSE
2	2.0	1.16	.18	.92	.81
7	2.4	1.90	.16	.77	.85
11	2.7	2.02	.14	.92	1.05
4	2.5	2.21	.17	.85	.88
1	2.2	2.36	.17	1.13	1.03
6	2.4	2.53	.16	1.18	1.24
3	2.7	2.76	.15	1.10	1.05
10	3.3	2.96	.15	1.13	1.13
9	2.5	3.05	.16	.79	.85
8	2.9	3.13	.16	.75	.92
5	2.8	3.47	.15	1.29	1.28

The rating exercise identified around 20 candidate scripts as being at the level of C1. This number was raised to thirty by the slight relaxing of the criteria for C1 to allow some high-scoring B2+ scripts to be judged as C1, particularly if they had been rated as C1 against aspects of the criteria. This relaxation of the standard was felt appropriate due to the stricter rating of some of the raters. The decision to relax the criteria can also be supported due to the difficulty clearly identifying between the B2+ band and C1. As section 3.7 has shown, in terms of discourse competence, C1 is an extension of the skills identified in the B2+ level and, as Weir points out (2005a), there is little guide as to how the actual abilities such as discourse competence, develop between levels.

One final point to note was that no scripts which were originally put forward for the B2 corpus were moved to C1 level and this was due to the fact that no raters allocated any of the B2 scripts up to the C1 level.

6.5.9 Ethical issues

All of the scripts used for the product analysis came were produced by Greek candidates in the ESB English Language Examinations and permission was gained from ESB for the research. All of the data was taken from completed sessions of the examinations and the awarding of grades had been completed. This ensured that the research process was entirely independent of and had no influence on the grades candidates received in the examinations. Once a paper had been typed up it was assigned a number, a level (B2 or C1) and the details of the age and gender of the candidate were recorded. With the candidate name, centre number and the year of administration removed there was no means of tracing any script back to its author. Texts and data were held securely in files using a password to protect the information.

6.6 Data analysis: Product

As stated at the start of this chapter, the purpose of largely quantitative strand of the study was to investigate the role of discourse competence in determining the levels of B2 and C1 level candidates of the ESB ESOL International Examinations. Metadiscourse markers were chosen as the feature for investigation for the reasons set out in 6.5.1 and the following research questions were set:

3. Is there a difference in the quantity of metadiscourse markers used by candidates of the ESB ESOL International Examinations at levels B2 and C1 of the CEFR?
4. What are the functions of the metadiscourse markers used by candidates at level B2 and C1?
5. To what extent do the metadiscourse functions employed by candidates at levels B2 and C1 match the predictions made by the

CEFR regarding the development of discourse competence in learners at these levels?

In order to investigate these questions quantitatively, it was necessary to apply a number of tests to the data collected.

6.6.1 *Coding of metadiscourse markers.*

The first step was to code the metadiscourse in the thirty texts at C1 and the thirty texts at B2. Hyland's 2005 metadiscourse scheme was used (see 4.6). However, before the analysis was begun the scripts were read through in order to identify any additional metadiscourse markers which might be considered. Table 6.8 shows the items which were added to Hyland's lists.

Table 6.8 *Lexical exponents added to Hyland's 2005 categories*

Metadiscourse Category	Added items
Endophoric Markers	In the last paragraph In the introduction
Sequencing	First and foremost Last but not least
Shift Topics	As regards to
Transition Markers	This means Which means On the one hand On the other side of the coin Every coin has two sides In order to
Hedges	In most Personally

The step of adding items to Hyland's list was necessary for two reasons. Firstly, while metadiscourse schemes are not necessarily open-ended in

terms of the linguistic exponents which they could contain, the metadiscourse functions could be carried out by more items than even Hyland's list of nearly 500 metadiscourse markers (2005, p. 218-224). Secondly, because the study involved texts produced by non-native speakers it was possible that the learners might use some phrases to carry out the functions that a native speaker would regard as incorrect or unusual. For example, the phrase 'on the other side of the coin' is one which strikes many native speakers as being odd, but it is one which frequently occurs in the writing of students from different countries (such as Greece and China). Possibly it is a direct translation from a phrase in their own L1 which is used as a transition marker. It was therefore necessary to ensure that such features were captured in the analysis.

Once the list of metadiscourse markers was complete, the 'find' function on Microsoft Word (Microsoft, 2013) was used for each marker systematically. It was originally the researcher's intention to use the Concordancer on Lextutor (Cobb, 2014) but trials with this found that it sometimes returned inaccurate results. By contrast Word allowed the researcher to find each instance of a particular word in the text and examine it in context to determine whether it was indeed carrying out a metadiscourse function.

Spreadsheets were used to record the metadiscourse markers by category. Three pieces of data were collected for each exponent. There were:

- (a) counts of the word/phrase according to Hyland's 2004 analysis;
- (b) counts of the word/phrase according to Hyland's 2005 analysis;
- (c) the number of incorrect uses of the word/phrase.

As set out in 4.5.1 of Chapter Four, Hyland's 2005 scheme had a more stringent test for metadiscourse function and there was a concern that this

would lead to some use of metadiscourse being ruled out of the analysis when in fact candidates might not have perceived a functional difference in the use. However, unlike some of the genres that Hyland considers (e.g. text books) the texts in this study are very short and most of the metadiscourse markers used fulfilled both the 2004 and 2005 criteria. In fact, the 2005 ‘test’ proved a useful aid in identifying metadiscourse function, particularly with modal verbs which had been a problematic group of words to analyse in the pilot (see 5.11).

Figure 6.2. *Excel (Microsoft, 2013) worksheet for analysis of metadiscourse markers (arranged alphabetically).*

Level	Script	Finally 2004	Finally 2005	Errors	First 2004	First 2005	Errors
C1	1	0	0	0	0	0	0
C1	2	0	0	0	0	0	0
C1	3	0	0	0	0	0	0
C1	4	1	0	0	0	0	0
C1	5	0	0	0	0	0	0
C1	6	0	0	0	0	0	0
C1	7	1	1	0	0	0	0

In the end, only the 2005 results and the errors were used to analyse the data.

6.6.2 Descriptive analysis of the data

The first stage of the analysis was to consider the descriptive statistics such as means and also to compare the means as part per-hundred. This step was important because Figures obtained from the raw data could be distorted by the C1 candidates having produced more text than the B2 candidates. For example, C1 candidates might appear to use more metadiscourse even if the overall percentage of metadiscourse in the text was lower.

6.6.3 Mann-Whitney U test for differences between B2 and C1

The Mann-Whitney U test was chosen as an alternative to a one-way ANOVA test. While Dornyei (2007, p. 227-228) identifies that **parametric**

tests such as the t-test or ANOVA have more statistical ‘power’, that is they are better able to identify statistically significant results, they are not suitable for use when the data does not display a normal distribution. The Mann-Whitney U test was selected because it is recommended for use when two nominal variables are being compared (Connolly, 2007, p, 176; Pallant, 2005, p.291). In order to use the test, hypotheses were generated for investigation. These are linked to the related research question in **Table 6.9.**

Hypothesis 1: There is a significant difference between the number of metadiscourse markers used by candidates at B2 and C1.

The null-hypothesis was the outcome expected for hypothesis 1 based on studies such as Burneikaite (2008) and Bax, Nataksuhara, & Waller (Forthcoming) as discussed in section 4.7.

Hypothesis 2: There is a difference in the proportion of interactive metadiscourse markers used between levels B2 and C1

Hypothesis 3: There is a difference in the proportion of interactional metadiscourse markers used between levels B2 and C1

Hypothesis 2 and 3 are intended to broadly investigate the functions of metadiscourse marker used by candidates at different levels. Interactive metadiscourse markers contain many of the metadiscourse markers which have been traditionally identified as being textual. Burneikaite (2008), Hyland & Tse (2004) and Bax, Nataksuhara & Waller (Forthcoming) have all identified that higher level candidates will use more interactional metadiscourse markers than lower-level ones.

Hypothesis 4: There are differences in the way individual categories of metadiscourse marker are used between B2 and C1.

For hypothesis 4, the Mann-Whitney Test was also used to look at differences in the use of categories of metadiscourse markers. It was anticipated that as well as there being a difference between the amount of interactive and interactional metadiscourse markers used there would be differences between the categories.

Hypothesis 5: There are differences in the way individual types of metadiscourse markers are used by B2 and C1 candidates.

This hypothesis aims to explore which exponents were used from each category of metadiscourse marker by level. The aim of this hypothesis was to identify whether higher level candidates did indeed abandon the simpler linguistic exponents (or lexical teddy bears (Hasselgren, 1994)) in favour of more complex metadiscourse markers.

Table 6.8. Summary of Research Questions and Hypotheses for the product analysis of metadiscourse markers.

Research Question (3-5)	Hypotheses	Test to be applied
RQ.3 Which metadiscourse markers are used by candidates of the ESB International ESOL Examinations at levels B2 and C1 of the CEFR?	H1: There is a significant difference between the number of metadiscourse markers used by candidates at B2 and C1.	Mann-Whitney U test
RQ.4 What are the functions of the metadiscourse markers used by candidates at level B2 and C1?	H2: There is a difference in the proportion of interactive metadiscourse markers used between levels B2 and C1	Mann-Whitney U test
	H3: There is a difference in the proportion of interactional metadiscourse markers used between levels B2, and C1	Mann-Whitney U test
	H4: There are differences in the way individual categories of metadiscourse marker are used between B2 and C1.	Mann-Whitney U test
RQ.5 To what extent do the metadiscourse functions employed by candidates at levels B2 and C1 match the predictions made by the CEFR regarding the development of discourse competence in learners at these levels?	H5: There are differences in the way individual types of metadiscourse markers are used by B2 and C1 candidates	Qualitative Analysis of corpus

It was also assumed based on previous studies (see 4.7) that higher level candidates would use metadiscourse markers more accurately in terms of function than at lower levels. In this stage of the research, the corpus was examined qualitatively in order to identify differences in the way candidates at the two levels use a range of metadiscourse markers. The decision to carry out a qualitative review was taken to lessen the impact of type one and type two errors due to the small size of the corpus and the use of the Bonferroni Correction (Pallant, 2005, p. 200).

This qualitative investigation comprised three stages. First of all, the numbers of metadiscourse markers in each category were manually inspected to identify differences in the numbers of each metadiscourse marker used. Concordancing software was used for the qualitative investigation, in this case the concordancing tool on the Compleat Lexical Tutor site (Cobb, 2014).

6.7 Conclusion

In this chapter the research methods and analysis tools have been set out. The process strand of the study is intended to investigate how discourse competence appears in the way participants at the B2 and C1 levels of the CEFR compose their scripts while the product strand aims to examine how the functions of discourse competence are used by writers at these two levels of proficiency. Chapter Seven will now explore the results from the product strand of the study and Chapter Eight shows the results from the process strand of the study. Both sets of findings will be discussed in Chapter Nine, bringing the two strands together.

Chapter Seven: Results from Process Strand of the Study

7.1 Introduction

Chapters Seven and Eight report the findings from the main study. This chapter will deal with the largely qualitative results collected from the verbal reports which relates to research questions one and two. In doing so the chapter will investigate the first aim of the study:

To what extent is **cognitive validity demonstrated** in the cognitive processes that candidates carry out while producing scripts in the English Speaking Board ESOL International Examinations?

Chapter Eight will report the results from the quantitative analysis of the products.

Chapter Seven will begin by recapping on how data from the verbal reports was analysed (7.2) and will then report the outcomes from the analysis. These outcomes will reported as descriptive statistics (7.3) and the qualitative data from the verbal reports and interviews. In section 7.4 data from the verbal reports is analysed using Independent T-Tests.

The implications of the results will be discussed in Chapter Nine where the results will be applied to the two research questions for this strand:

1. What cognitive processes do candidates at levels B2 and C1 appear to employ when composing timed essay tasks?
2. To what extent does the timed essay question format used in the ESB ESOL International Examinations elicit the cognitive processing that models predict at levels B2 and C1?

7.2 Recap on transcription & NVivo analysis

As was set out in 6.4 coding of the transcripts was carried out through an initial exercise carried out on paper. This first stage was intended to allow data to emerge without the pre-application of a theoretical model (Dornyei, 2007, p. 251). The resulting categories were then tested with a second rater and adjustments made resulting in the scheme which was then used to code the transcripts in NVivo (QSR International, 2014). NVivo was used because it allows the researcher to view the sections coded by category (or “nodes” as they are termed in the NVivo approach) in order to examine what participants have actually said at each point. The programme can also generate graphs and Tables to illustrate the coding of the data. This data is presented in 7.3 below.

Once NVivo (QSR International, 2014) coding had taken place the data was compiled in Excel (Microsoft, 2013) and imported into SPSS (IBM Corp, 2013) in order to carry out quantitative analysis of the results in order to test for statistical significance. This was done in order to see whether some of the patterns observed in the NVivo output could be applied to a wider population. Connolly (2007, p. 7) makes the point that significance is not necessarily always the most important outcome from data; the emergence of different patterns can also be revealing but it was hoped that the data would highlight behaviour on the part of the participants which might suggest differences in the way the wider test-taking populations at B2 and C1 respond to timed writing tasks.

7.3 Descriptive statistics based on quantitative analysis of transcripts

7.3.1 Data from the verbal reports

NVivo (QSR International, 2014) was used to explore the coded transcripts. Table 7.1 presents the number of nodes coded proportionally between B2 and C1.

The raw results show that in terms of the proportion of nodes coded at B2 and C1 there are differences in the processes the participants carried out. While all the identified nodes were employed by at least one candidate at each level, the data shows that there were more instances of each node among the C1 candidates and that more C1 candidates made use of the processes than the B2 candidates did.

Table 7.1. *Number of nodes coded in B2 and C1 transcripts.*

B2			C1		
Nodes	Number of coding references	Number of scripts in which codes were found	Nodes	Number of coding references	Number of scripts in which codes were found
Consider Audience	1	1	Consider Audience	4	3
Generating Content	26	5	Generating Content	31	6
Task Assessment	10	6	Task Assessment	12	6
Word Count	2	1	Word Count	2	2
Immediate Planning	61	6	Immediate Planning	84	6
Linking paragraphs	1	1	Linking paragraphs	3	3
Summarising Content	5	4	Summarising Content	8	4
Monitoring (Unspecified Purpose)	1	1	Monitoring (Unspecified Purpose)	3	3
Monitoring Content	5	3	Monitoring Content	7	3
Monitoring Language	4	3	Monitoring Language	8	4
Text Level Organisation	2	1	Text Level Organisation	10	4
Revising	5	2	Revising	13	6
Stuck	1	1	Stuck	3	1
Time Pressure	1	1	Time Pressure	4	2
Searching for Lexis	1	1	Searching for Lexis	14	4

For both groups the nodes of Generating Content and Immediate Planning were the most utilised nodes. We can also see that although four C1 participants carried out the function of Searching for Lexis while composing their texts, and did so 14 times between them, only one B2 candidate referred to this function on a single occasion.

However, these Figures are descriptive and fail to take into account a number of issues. Firstly, the C1 participants tended to write for longer in the verbal reports. Table 7.2 shows the number of interventions by participant and the mean number of interventions in each verbal report (by intervention we mean the number of two-minute intervals when participants were stopped and asked to report on what they had been thinking about and doing). Despite that fact that both the B2 and the C1 participants were given the same amount of time for writing (up to 45 minutes of actual writing time as they would have in the actual exam) it is unsurprising that the C1s should have written for longer, their writing task required more words (250-280 words versus the 150-180 words at B2). Therefore it is possible that the higher number of instances for each node could simply be as a consequence of the fact that the C1 participants were writing for a more extended period of time and therefore reported more. Even without this proviso, descriptive statistics are extremely limited as a basis from which to generalise about the behaviour of a wider population (Dornyei, 2007, p. 209).

Table 7.1 also shows the how many participants used each of the processes in the B2 group and C1 group. Again, the proviso must be noted that the C1 participants by taking longer to produce their scripts had more opportunity to describe processes but the information can still provide initial insights into the behaviour of the participants in the study.

Table 7.2. *Mean number of interventions by participant.*

B2		C1	
Participant	Number of interventions	Participant	Number of interventions
4	10	1	11
5	12	2	22
7	11	3	22
8	12	4	19
9	12	11	13
10	15	12	18
Mean # of interventions at B2		Mean # of interventions at C1	
12.00		17.5	

What is particularly striking about the frequency information presented in Table 7.1 is that the C1 participants appear to make use of or at least describe a wider range of processes while they are writing. For example, Text Level Organisation was only mentioned explicitly by one of the B2 participants when they said:

Extract 7.1

(a)

I tried to underline key words in order to have a plan in my mind in order to write introduction and then er write the <\$=> the whole writing <\$E> pause <\\$E> and I started doing the plan and I will put <\$E> pause <\\$E> the introduction first so I'm taking notes here in order to write there.

(Participant 8, B2 <02:09>).

(b)

<\$2> I first er wr= wrote er why th= they were living er more erm so I can tell the benefits <\$E> pause <\\$E> er in the first paragraph and in the second I'm going to write the negative part and the conclusion is going to be one together.

(Participant 8, B2 <11:31>).

This is in contrast to four participants at C1 who referred to Text-Level Organisation and on the face of it this seems to signal something about how the B2 and C1 participants in the study approach the writing task. The question of whether the differences between the groups is statistically significant and can be generalised to the wider test-taking population is addressed in 7.4 below.

7.3.2 *Post-interview data*

The post-interviews were mainly used to explore what participants had reported during the verbal reports and as discussed in 6.3.8, three additional questions were asked of which i and ii were used to inform the coding process. The three post-interview questions were:

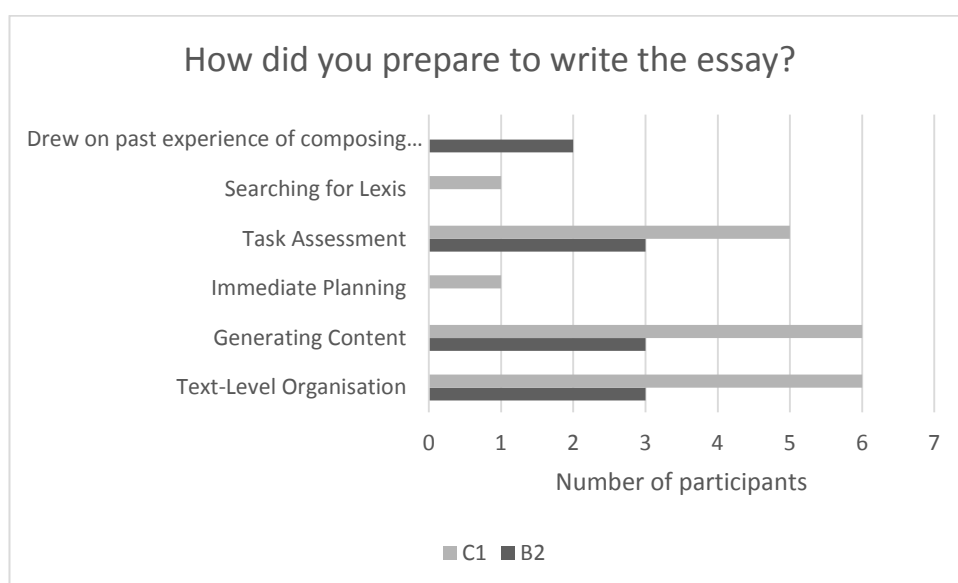
- i How did you prepare to write the essay?
- ii What did you think about after you had completed the piece of writing?
- iii When you are writing an essay like this in an exam, who do you think you are writing for?

It is important to note that as set out in 6.3.8, these questions were not strictly part of the verbal report and so the results must be treated with caution. The participants could potentially draw on long-term memory to answer the questions rather than just describing what they had done.

Figure 7.1 shows the responses to question i and what participants had identified that they attended to when they prepared to write the essay.

Both the B2 and the C1 participants mentioned carrying out activities which are wider than Macroplanning, which to some extent emphasises Field's observation that the phases are not linear (2004). The C1 participants again reported more processes and placed more emphasis on consideration of the task, drawing together resources (such as content and lexis) and planning out the text. Fewer of the B2 participants talked about task assessment or pre-planning but two of them reported that they had drawn on their memories of producing previous essays which they deemed 'similar' to the task they had been given (see extracts 7.2 a and b below).

Figure 7.1. *B2 and C1 participant responses to the post-interview question 'How did you prepare to write the essay?'*



Extract 7.2

(a)

<\$2> No <\$E> pause <\\$E> writing many essays in the whole year so <\$E> pause <\\$E> it was not necessary for me to <\$E> pause <\\$E> practice and that or prepare.
<\$E> prompt at 26:44 <\\$E> <\$1> so what did you think about when you were starting the essay?

<\$2> Er It was something familiar for me so <\$E> pause <\\$E> I was not nervous.

(Participant 4, B2 <26:17>)

(b)

<\$2> I started thinking of <\$E> pause <\\$E> things I can write and of course I used my background knowledge my experience because <\$E> pause <\\$E> the topics we are given to write are <\$E> pause <\\$E> very co= co <\$=> are common <\$E> pause <\\$E> and they are given to us many times so we remember what we've written I hope. <\$E> pause 3 seconds <\\$E> At least me.

(Participant 7, B2 <36:58>).

Figure 7.2. B2 and C1 participants' post-interview responses to the question *What did you think about after you had completed the piece of writing?*

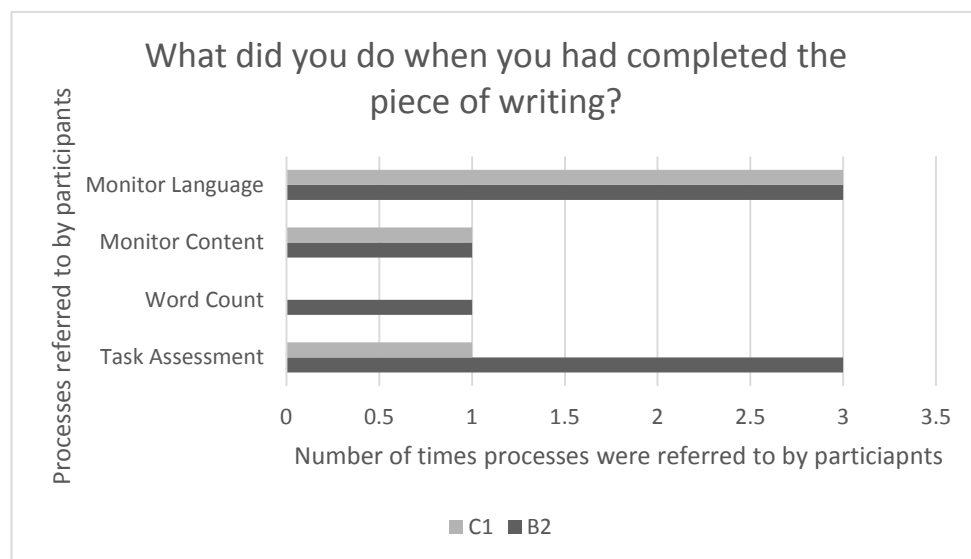


Figure 7.2 shows the responses to the question about what the participants had thought about at the end of the piece of writing. Both groups tended to monitor the language and content. B2 participants commented on how they felt about the piece of writing (being 'happy' with it or how 'easy' they thought the task was) which has been included in the

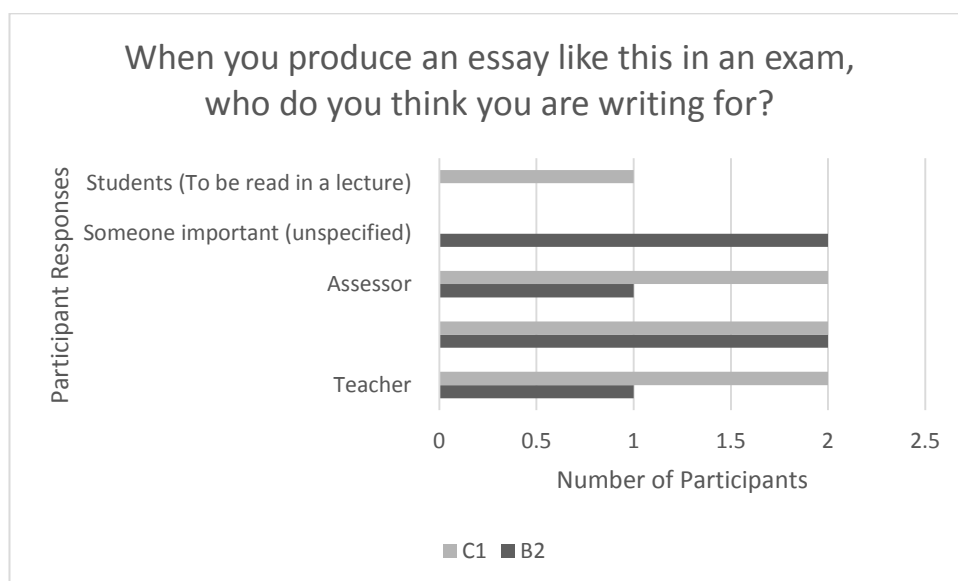
task assessment category while the C1 participant appeared to be more explicitly thinking about the demands of the question:

Extract 7.3

Erm I'm thinking if I have explained everything that is asked
(Participant 11, C1 <36:35>)

The question asked at the end of the interview related to the intended audience for the task. This question was deliberately asked due to the issue of the lack of explicit audience identified in the writing task (see section 6.5.4) in order to see what impact this had on participants. Figure 7.3 reports the results from this question.

Figure 7.3. B2 and C1 participants' responses to the post-interview question *When you produce an essay like this in an exam, who do you think you are writing for?*



The results show that the C1 participants had a clearer picture of the person or persons to whom they were writing: either an assessor, a teacher or an audience in a more formal context. Two of the B2 participants were unsure beyond the notion of it being “someone important”. However, all of the participants brought to the piece of writing the notion that the readership was more formal and so demanded more

formality in terms of style and increased the need for accurate language use.

7.4 Inferential statistics based on quantitative analysis

As set out in 7.3 above, descriptive analysis of the B2 and C1 transcripts identified that in terms of the coding the C1 participants used more of the following categories:

- Consider Audience
- Linking Paragraphs
- Text Level Organisation
- Revision

By contrast, the B2 participants did more Immediate Planning despite having fewer interventions to report. The use of inferential statistics was required to check for statistical significance in these differences.

The first step was to ensure that the data from the verbal reports was normally distributed as the independent t-test is not available when data does not have normal distribution (Connolly, 2007, p. 201). As recommended by Connolly, a one-sample Kolmogorov-Smirnov test was run on the data which indicated that all categories were normally distributed thus allowing the independent t-test to be used.

A two-tailed independent t-test was used to compare the B2 and C1 participants for each category. The two-tailed test was chosen following Connolly (2007) and Dornyei's advice to do so when unsure of the outcomes (2007, p. 210-211), since the differences in the numbers of interventions between the two groups made it difficult to predict which groups would have higher scores. The t-test was used to examine whether there was a relationship between the level of the candidates and the processes they reported. The results are presented in Table 7.3

below. Only two categories of reported behaviour had statistical significance: 'searching for lexis' and 'revising'. Text-Level Organisation was close in terms of significance with B2 ($M = 0.33$, $SD = 8.16$) and C1 ($M = 1.67$, $SD = 1.37$), $t(10) = -2.052$, $p=0.06$. The effect size was calculated using Pallant's formula (2005, p. 201) and the Figures for effect size in Table 7.3 suggest that the results for searching for lexis and revising can be considered large effects in that they are above 0.14.

Table 7.3. *Independent samples t-test for all categories in the verbal reports (N=12).*

Category	Level	Mean	Std. Deviation	d	t	Sig	Effect size ^a
Task Assessment				10	-0.79	.45	0.14
	B2	1.67	.82				
	C1	2.00	.63				
Generating Content				7.05	-.48	.64	0.09
	B2	4.17	3.87				
	C1	5.00	1.79				
Word Count				10	0.00	1.00	0.00
	B2	.33	.82				
	C1	.33	.52				
Consider Audience				10	-1.34	.21	0.21
	B2	.17	.41				
	C1	.67	.82				
Text Level				10	-2.05	.06	0.29
Organisation	B2	.33	.82				
	C1	1.67	1.37				
Immediate Planning				5.898	-1.02	.33	0.17
	B2	9.67	1.97				
	C1	12.50	6.53				
Linking Paragraphs				10	-1.19	.26	0.19
	B2	.17	.41				
	C1	.50	.55				
Summarising Content				10	-.78	.45	0.14
	B2	.83	.75				
	C1	1.33	1.37				
Searching for Lexis				5.39	-2.52	.03*	0.34
	B2	.17	.41				
	C1	2.33	2.07				

Monitoring for				10	-1.19	.26	0.19
Unspecified Purpose	B2	.17	.41				
	C1	.50	.55				
Monitoring Content				10	-.222	.83	0.04
	B2	1.00	1.26				
	C1	1.17	1.33				
Monitoring Language				7.83	-1.45	.18	0.22
	B2	.50	.55				
	C1	1.17	.98				
Revising				10	-2.18	.05*	0.30
	B2	.67	1.21				
	C1	2.17	1.17				
Time Pressure				10	-.45	.66	0.08
	B2	.17	.40				
	C1	.33	.82				
Stuck				10	-.63	.54	0.11
	B2	.17	.40				
	C1	.50	1.23				

* $p < 0.05$

^a Eta squared

The results gained from the independent t-test by category provided two areas of statistical significance but the researcher also wished to consider the results as larger processes using Field's cognitive phases (2004). For this analysis, categories were combined as set out in Chapter Six (6.4.4 and Appendix 10). Categories from outside Field's phases (i.e. the categories of Time Pressure and Stuck) were excluded since these had already been examined in the analysis in Table 7.3 above. Categories such as Organisation, Translating and Revising were also not expected to change as they comprised only one sub-component.

As the data had been established as having normal distribution, an independent two-tailed t-test was used to analyse the data. The results are in Table 7.4. However, despite combining the categories using Field's cognitive phases (2004) there were no changes to the results and no other areas demonstrated statistical significance.

It was then decided to look at the question of when different processes had been used during the writing process. The researcher hoped to find out when the participants had referenced using particular phases during the essay writing to see whether there were differences between the behaviour of the B2 and the C1 students. As set out above, the descriptive statistics and the post-interview questions suggested that there were difference in the way that the two groups went about composing their scripts (see Figures 7.2 and 7.3).

Table 7.4. *Independent samples t-test for Field's cognitive phases used in the verbal reports (N=12).*

Phase	Std.		d	t	Sig	Effect Size ^a
	Mean	Deviation				
Macroplanning			10	-.91	.38	0.15
	B2	6.33	3.83			
	C1	8.00	2.28			
Organisation			10	-2.05	.07	0.29
	B2	.33	.81			
	C1	1.67	1.36			
Microplanning			5.67	-1.15	.28	0.19
	B2	10.67	1.96			
	C1	14.33	7.52			
Translation			5.39	-2.52	.03*	0.34
	B2	.17	.40			
	C1	2.33	2.06			
Monitoring			10	-1.30	.22	0.21
	B2	1.67	1.63			
	C1	2.83	1.47			
Revising			10	-2.18	.05*	0.30
	B2	.67	1.21			
	C1	2.17	1.16			

* p < 0.05

^a Eta squared

As shown in Table 7.2, the participants took different amounts of time to produce their scripts, even within levels. The mean was taken and with the C1 mean at 17.5 it was rounded up to 18. This was also felt to be convenient as it would allow the processes to be divided into three comparable periods: start of writing, mid-writing and the end of the process. Each participant's data was broken up into three parts with any 'leftover' interventions added to the final turn (e.g. in the case of candidates 2 and 3 who both had 22 interventions, the first third consisted of turns 1-7, the second third was turns 8-14 and the final third was turns 15-22).

No statistically significant differences were found in the first third and unsurprisingly there were no instances of 'Summarising Content, Monitoring' or 'Revision' in this section since the participants had just started writing. Likewise, no statistically significant results were found for the second third. In the final third, a significant difference was found for Macroplanning between the B2s ($M=1.83$, $SD=1.17$) and C1s ($M=.50$, $SD=.55$), $t(10) = 2.53$, $p<.05$).

7.5 Qualitative comments by category

The data from each category which was coded in NVivo (QSR International, 2014) was also examined for themes which appeared in the comments that participants made.

7.5.1 Task Assessment

Both B2 and C1 participants commented on Task Assessment. The topic was frequently mentioned including thoughts about what information might be relevant to it. This was done both at the beginning (as in extract 7.4 a and b) and linked to other processes such as Generating Content and in some cases at the end (extract 7.4 c) when it was part of an 'evaluation' comment by the learner either on the suitability of what they had written or on the task as a whole (extract 7.4 d).

Extract 7.4

(a)

I read very well er all the task er I try to understand exactly what it wants

(Participant 1, C1 <02:04>)

(b)

Firstly I read the subject and then I <\$E> pause <\\$E> always do erm brain storming schedule

(Participant 2, C1 <02:40>)

(c)

I am <\$E> pause <\\$E> being thinking if I have to write <\$=> I think it's good or bad <\$E> pause <\\$E> for the topic.

(Participant 10, B2 <38:59>)

(d)

I think that it was more easier than <\$=> oh no it was easier than I thought <\$E> pause <\\$E> I was waiting it to be something very difficult <\$E> pause <\\$E> it is <\$=> it was like the writing composition that we write here in this school so it was easy. Yes I believe I did well.

(Participant 5, B2 <31.08>)

The task was also referred to by participants in order to decide whether content or techniques would be appropriate. For example, Participant 7 (B2) asked "Does it matter if I use a rhetorical question?" (<24:43>).

7.5.2 *Generating Content*

As reported in the post-interview data, the C1 participants stated that their pre-planning involved the brainstorming of ideas and Table 7.5 which

shows the mean scores for instances of Generating Content reported for both groups in each third of the time it took participants to compose their essays supports this. What is interesting is that while C1 instances of Generating Content decline, suggesting that these writers have done more to pre-plan what they will write, the B2 Figure increases in the second third.

Table 7.5. *Means for Generating Content across writing time (N=12).*

	Mean during first third of writing time	Mean during second third of writing time	Mean during final third of writing time
B2	1.16	2.16	.83
C1	2.83	1.83	.33

Generating content at the start of writing was sometimes connected to the assessment of the task as in extract 7.4 b above and 7.5 a and b below.

Extracts 7.5

(a)

Er I thought about erm what kind of essay I'm going to write. And erm then I thought about the arguments and erm what I was asked to write

(Participant 11, C1 <02:40>).

(b)

...I start to think the erm <\$=> what er the general erm topic.

<\$1> Mmhmm.

<\$2> And I quite copy from the erm instructed <\$E> laughs <\\$E> so this is what I've been doing up to here and I've meanwhile I've been thinking of the ideas. I've been brainstorming on what I can write.

(Participant 3, C1 <02:35>)

Later references to Generating Content appeared to be either seeking out additional examples or else attempting to come up with ideas for the essay. The B2 participants in particular appeared to be Generating Content as they went through the essay as in the case of 7.6 a and b which were reported when the candidates were into the second third of their writing time.

Extract 7.6

(a)

<\$2> I'm trying to find some positive and negative impacts of this increase of longevity <\$E> pause<\\$E> and so I'm trying to think because <\$E> pause <\\$E> I haven't thought of it before. <\$E> laugh<\\$E>so it is my first time thinking about it. And I'm trying to find some impacts positive and negative.

(Participant 4, B2 <11:10>)

(b)

I was erm trying to <\$=> to think some erm <\$=> some reasons oh no <\$=> some positive impacts but I don't erm <\$=> I have not think <\$=> have not thought a lot of <\$E> pause 2.5 seconds <\\$E> er but I have remembered that er people who live at <\$=> lived years ago er they don't live a lot they <\$E> pause <\\$E> erm live er in fifty years <\$E> pause <\\$E> like that.

(Participant 5, B2 <16:03>)

What is striking about the B2 participant comments above is that they occurred close to half-way or more through the writing process but the participants were still attempting to come up with ideas for their essay. This suggests that although the participants may have had an overall plan for their piece of writing in terms of writing advantages and then disadvantages, they had not considered what would constitute the points of the essay, nor how they would link. This represents a 'hand to mouth'

approach to essay writing in that content was generated and then immediately written down.

Other instances of Generating Content were reported when participants were seeking examples to further illustrate their arguments or when ideas came to them during writing. Extract 7.7a reports an instance where a participant while working on the later parts of the essay is struck by an idea to add something to the first paragraph. Participant 3 mentions 'catching' an idea (<35:11>) suggesting that the act of writing may have sparked a connection that she had not previously considered. Most of the references to coming up with examples, that is illustrations to support what has already been written (as in extract 7.7 b) are in the C1 scripts while the B2s seem to be actually in the process of coming up with the main ideas for their writing, even quite late in the process as in the extracts 7.8.

Extracts 7.7

(a)

I thought of the first paragraph <\$E> pause <\\$E> and <\$E> pause <\\$E> I came up with the idea of adding something
(Participant 6, C1 <39:40>)

(b)

and erm I'm trying to find an example in order to explain it.
(Participant 11, C1 <12:28>)

Extracts 7.8

(a)

<\$2> Okay. I finished with the first topic. I wanted to write about the medicine and now I just started thinking of the next thing that helps our <\$=> us to have a longer life expectancy. <\$E> pause <\\$E> I haven't thought of it yet.
(Participant 7, B2 <10.02>)

(b)

<\$2> Er I found the topic about my third paragraph.

(Participant 10, B2 <23:46>)

7.5.3 Consider Audience and Word Count

There were very few instances of the Consider Audience category in the verbal reports. Those statements reported by C1 candidates tended to be about the effect on the reader (see extracts 7.9 a, b and c).

Extracts 7.9

(a)

\$2> I'm trying to think er the idea that er <\$E> pause <\\$E> once someone reads it er it makes a difference in all the other er essays I will read+

<\$1> Hmm

<\$2> +I think that this always improves er the writing.

(Participant 1, C1 <15:28>)

(b)

<\$2> Erm I'm thinking what to write in the second paragraph and how I will start erm to <\$E> pause <\\$E> prepare who is going to read the essay about what I am writing in this paragraph.

(Participant 11, C1 <07:29>).

(c)

I <\$E> pause <\$E> always try to erm make it easy for the reader to see how my thoughts are linked

(Participant 2 post-interview, C1 <01:06:51>)

There was only one B2-level comment which did mention the intended audience but it occurred when the candidate was asked about their remarks in the post interview (extract 7.10).

Extract 7.10

I want to analyse bits so the person who reads it will understand bit why I wrote that as a disadvantage.

(Participant 10, B2 <44:17>)

Word Count was also only mentioned by three participants. Participants 2 and 3 at the C1 level worried that they might be writing too much while Participant 8 (B2) counted their words to see how much they had written.

7.5.4 Text-Level Organisation

Text-Level Organisation was mainly used by C1 participants who particularly referred to the category in the first third of their writing time. References were often in the context of planning the piece of work as in extracts 7.11 a and b.

Table 7.6. Means for Text-Level Organisation across writing time (N=12).

	Mean during first third of writing time	Mean during second third of writing time	Mean during final third of writing time
B2	.17	.17	0.00
C1	1.3	.17	.17

Extracts 7.11

(a)

in my mind I try to separate all the paragraphs to see how many words will I use. Now I know exactly what to write in each paragraph for example.

(Participant 1, C1 <02:04>)

(b)

I've been thinking about erm making a plan first outlining er er I actually wanted to write about three points but this is about erm er

why people are living much longer nowadays and the other points are the negative impacts and the other point is the positive.

(Participant 12, C1 <02:17>)

References to Text-Level Organisation later on in the writing task were amendments to the plan that the participant had already described (extract 7.12).

Extract 7.12

I've just finished the introduction and er I'm thinking about writing the second paragraph. Erm I'm thinking to separate erm er two <\$=> to write two paragraphs. In one paragraph I'm going to write about erm <\$=> I will explain why people's life expectancy has <\$E> pause <\\$E> risen and erm I'm thinking about erm writing the positive and negative impacts <\$E> pause <\\$E> in one paragraph <\$=> the third one

(Participant 12, C1 <10:00>).

Field notes made by the researcher indicate that only five of the twelve participants made notes before they started writing. Table 7.7 provides the details.

The Figures in Table 7.7 suggest that Text-Level Organisation may be under-represented in the verbal report data. It is also worth noting that the B2 participants tended to write down key words from the questions in a list rather than generating ideas so it is not entirely clear that these instances are distinct from the Task Assessment category. The other point that is worth considering is that participant 2 spent a considerable amount of time (almost the first third of her time) making notes some of which consisted of complete chunks of text which were later copied into the essay ("I tried to write erm <\$=> copy my notes quickly" <56:50>). In this instance, planning appears to have comprised a rough draft of the essay which could then be

amended and adapted as it was incorporated into the final product. The remaining three C1 participants, like the four B2 participants began writing right from the beginning of the allocated essay production time.

Table 7.7. *Instances of note-taking prior to writing by participants.*

Level of Participant	Participant number	Time spent making notes
B2	8	0 – 4 mins
B2	9	0 – 4 mins
C1	2	0 – 14 mins
C1	11	0 – 2 mins
C1	12	0 – 4 mins

7.5.5 Immediate Planning

As reported in Chapter Six, Immediate Planning ended up subsuming a number of other categories identified in the initial coding exercise due to the difficulty in distinguishing between them. Unsurprisingly, the resulting category showed the most instances of use and was used by all twelve participants. Table 7.1 show that the B2 participants did more Immediate Planning despite taking less time to produce their essays but the C1 mean was higher ($M=12.50$) than the B2 participants ($M=9.67$) when the independent t-test was applied. Immediate Planning did not return a statistically significant difference. Means for the two groups of participants were also very close when examined across their writing time (Table 7.8).

Table 7.8. *Means for Immediate Planning across writing time (N=12).*

	Mean during first third of writing time	Mean during second third of writing time	Mean during final third of writing time
B2	3.00	3.66	3.00
C1	3.50	4.33	4.66

The majority of the comments in this category were to do with planning the piece of text which the participant was producing. Table 7.9 provides a frequency list of the most frequent words (pause and 'erm' which were the two most frequent items have been removed). This shows that most comments made during the verbal reports were to do with the paragraph being produced, or else focussed on the introduction or conclusion. Participants were also concerned with the 'positive' or 'negative' impacts which they were writing but also discussed the 'argument' they were making.

Table 7.9. *Frequency of lexis used by participants in the category 'Immediate Planning'.*

Rank	Word	Count	Weighted Percentage (%)
1	paragraph	41	1.93
2	write	35	1.65
3	first	34	1.60
4	writing	30	1.41
5	now	27	1.27
6	thinking	26	1.22
7	second	23	1.08
8	trying	23	1.08
9	negative	22	1.03
10	people	22	1.03

11	conclusion	19	0.89
12	going	18	0.85
13	positive	18	0.85
14	wrote	18	0.85
15	seconds	17	0.80
16	just	16	0.75
17	thought	16	0.75
18	finished	13	0.61
19	impacts	13	0.61
20	longer	13	0.61
21	argument	12	0.56
22	okay	12	0.56
23	find	11	0.52

7.5.6 Linking Paragraphs and Summarising Content

Both of these areas were identified as being part of Microplanning for the analysis but neither category generated statistically significant results. C1 participants mentioned linking paragraphs three times (extracts 7.18 below).

Extracts 7.18

(a)

Okay I tried hard to connect my first paragraph with the second so <\$E> pause <\\$E> I do not have er mix up the er meanings.

(Participant 2, C1 <41:09>)

(b)

I'm still thinking of how to write the first paragraph in order for me to be more easily=easier. To continue to the second paragraph

(Participant 6, C1 <04:54>)

(c)

I was closing the second paragraph and as I was closing it <\$=> as I was writing the words I was thinking of the paragraph of the first sentence of the second paragraph.

(Post interview - Participant 3, C1 <59:00>)

The comments from the B2 participant (Participant 8, B2 <16:45>) were to do with contrasting an idea from the previous paragraph.

Comments made during the verbal reports regarding summarising were all connected to writing the conclusion of the essay and with the exception of one mention they all occurred in the final third of writing time. Table 7.10 lists the most frequent words in the Summarising Content category (with pause, the number one word, removed). It shows that most of what the participants reported was that they were drawing together their arguments for the conclusions.

Table 7.10. *Most frequent words in the Summarising Content category.*

Word	Count	Weighted Percentage (%)
arguments	4	3.51
everything	3	2.63
paragraph	3	2.63

thinking	3	2.63
tried	3	2.63
trying	3	2.63
write	3	2.63
composition	2	1.75
conclusion	2	1.75
end	2	1.75
mention	2	1.75
sum	2	1.75
summarise	2	1.75
summary	2	1.75
things	2	1.75

7.5.7 Searching for Lexis

This category was one which proved to show statistically significant differences with only one instance of it being reported by B2 participants while four of the C1 participants reported doing this. B2 participant 7 (<07:25>) reported being stuck while thinking of a word but the C1 reports were more to do with selecting what they considered to be the right word for the task as shown in the extracts in 7.14.

Extracts 7.14

(a)

I'm trying to find some good words and not just thinking of boring words to put in er my es= essay

(Participant 1, C1 <05:13>)

(b)

<\$1> ...you said you were looking for good words and you were looking for nice words. What did you mean by that?
<\$2> Erm some words maybe for us that it's not our <\$=> our language but some words er are <\$E> pause <\\$E> more <\$E> pause <\\$E> um how can I say it er <\$E> pause 2.5 seconds<\\$E> when you read them are more <\$E> pause 3 seconds <\\$E> they stay in your mind better than other words....er the reader who reads my essay er he would like to see some words that er he will understand that er someone who is not his language <\$=> it is difficult to put these words in the essay.

(Post-interview – Participant 1, C1 <31:00>)

(c)

I'm <\$=> I'm trying to find word er to <\$E> pause <\\$E> express that the quality of life now is <\$E> pause <\\$E> better or something like that.

(Participant 2, C1 <35:18>).

(d)

so I faced again the problem of er finding er another word for tradition

(Participant 2, C1 <43:50>)

(e)

and I'm trying to write it er appropriately

(Participant 3, C1 <05:22>)

(f)

I'm actually trying to enrich it with er good vocabulary

(Participant 3, C1 <18:06>)

(g)

I was thinking about the appropriate= appropriate right <\$=>
<\$E> pause <\\$E> the appropriate word to use <\$E> pause
<\\$E> in my sentence

(Participant 6, C1 <34:02>)

Use of NVivo's (QSR International, 2014) frequency report (Table 7.11) shows that the words 'appropriate' and 'right' were both in the top eight and neither of these words were used by the B2 participant. This suggests that the C1 participants were concerned not only with finding a word to express their ideas but what they saw as being the most appropriate word for the task.

Table 7.11. *Frequency of words used for Searching for Lexis.*

Rank	Word	Count
1	word	11
2	trying	7
3	find	5
4	words	5
5	appropriate	4
6	put	3
7	right	3
8	thinking	3

7.5.8 Monitoring (Unspecified, Content and Language)

Monitoring covered three distinct categories which were identified in the initial coding and were maintained for the final coding of the data. The Unspecified category had four instances, three of which were from C1 participants. These instances consisted of statements that the participants had been re-reading or proofreading without any further indication either in the verbal report or in the post-interview as to what they had been looking at specifically.

Monitoring Content was mentioned by three of the B2 and three of the C1 participants and tended to occur in the middle and towards the end of the writing time (see Table 7.12) for the C1s while there was one instance of a B2 participant Monitoring for Content in the first third of time while the rest occurred in the final third.

The B2 comments tended to be evaluative and were made when the participant had reached the end of composing their essay. Extract 7.15 provide examples of these.

Table 7.12. *Means for Monitoring Content across writing time (N=12).*

	Mean during first third of writing time	Mean during second third of writing time	Mean during final third of writing time
B2	.16	.00	.17
C1	.00	.50	.33

Extracts 7.15

(a)

I am <\$E> pause <\\$E> being thinking if I have to write
<\$=> I think it's good or bad <\$E> pause <\\$E> for the topic.
(Participant 10, B2 <38:59>)

(b)

and I think that all the ideas I mentioned are enough for the
essay
(Participant 7, B2 <30:54>)

The comments made by one of the B2 participants (see the extracts in 7.16) were similar to those made by the C1 participants in that Monitoring for Content included comparing what had been produced with the plan.

Extract 7.16

(a)

I am reading the whole thing. In order to <\$E> pause <\\$E> <\$=> to ensure that I <\$=> I <\$=> what I'm writing has sense. And I'm checking my plan to see if I'm <\$E> pause <\\$E> in not only the words but also my ideas if are connected.

(Participant 8, B2 <21:56>)

(b)

I'm thinking about erm going back to the topic or my notes to see if I have forgotten something again to complete it.

(Participant 11, C1 <27:40>)

Monitoring for Language as a category started earlier for the C1 participants than it did for the B2s (See Table 7.13) suggesting that this was more of an ongoing process for these participants while for the B2 participants it was more of a summative activity.

Table 7.13. *Means for Monitoring Language across writing time (N=12).*

	Mean during first third of writing time	Mean during second third of writing time	Mean during final third of writing time
B2	.00	.00	.83
C1	.00	.33	.66

C1 participants also stated that they were not only monitoring for errors but in some cases seeking better ways of expressing their ideas or for stylistic purposes (see Extracts 7.17).

Extracts 7.17

(a)

I <\$E> pause <\\$E> corrected erm <\$E> pause <\\$E> a phrase here because it would be the same in right there
(Participant 2, C1 <59:17>)

(b)

I've been correcting some of my <\$=> some of the words I've written <\$E> pause <\\$E> to erm make better my vocabulary.

(Participant 3, C1 <32:44>)

7.5.9 Revising

As shown in Table 7.14, the C1 participants carried out more revision in the verbal reports than the B2 participants and although both groups started in the second third of the writing time, more of the C1 participants started earlier. For the B2 participants, revising followed the monitoring of their essay after they had finished while for the C1 participants it appears to have been more of an ongoing process during composition.

Table 7.14. *Means for Revising across writing time (N=12).*

	Mean during first third of writing time	Mean during second third of writing time	Mean during final third of writing time
B2	.00	.16	.50
C1	.00	.50	1.67

Comments by the C1 participants appeared to be fairly evenly divided between revisions to language and revisions to content. Both categories appeared throughout with content being assessed as it was produced (see extract 7.17) as well as at the end of the composition process.

Extract 7.17

<\$1> Erm <\$=> Right you got to the top of here <\$E>
pause <\\$E> of page two and there was a very long
pause
<\$2> Yes
<\$1> +about that point <\$E> pause <\\$E> a lot of pausing
and a lot of rubbing out
<\$2> Yes
<\$E> prompt at 55:27 <\\$E> <\$1> +then writing again
<\$E> pause <\\$E> could you remember what you were
thinking about at that point? <\$E> inaudible <\\$E>
<\$2> Erm I was brainstorming ideas but then I thought it
was not correct <\$=> incorrect because it was a bit out of
the topic <\$E> pause <\\$E> that's why I was erasing.
<\$1> Okay erm <\$=> and then you crossed out some
stuff down here
<\$2> Yes
<\$1> +you said "So because I thought what I had been
writing was incorrect" <\$E> pause <\\$E> what did you
mean by that?
<\$2> Erm <\$=> again it was irrelevant to the task given
(Post-interview - Participant 6, C1 <55:27>)

7.6 Conclusion for results from the investigation from the process strand of the study

The results show that the B2 and C1 participants employed all of the categories which were coded in the study. As shown in Table 7.1, in terms of numbers the C1 participants employed each category more frequently. Only the differences between two categories showed statistical significance overall: Searching for Lexis and Revising. In both of these categories the mean for the C1s was higher indicating that the higher group of participants made greater use of these categories.

When the data was examined according to time by breaking production into thirds, the only statistically significant difference was in macroplanning in the final third of writing time.

Examination of the verbal reports and post-interviews indicates that there are some differences in what the B2 and C1 participants attended to when writing. According to the post-interviews (Figure 7.2) C1 participants appeared to put more emphasis on question analysis and planning the response (Generating Content and Text-Level Organisation) before writing while some B2 participants drew on their experience of writing essays that they considered to be similar. The C1 participants also appeared to have a clearer idea of the audience for their essay (Figure 7.4). B2 participants appeared to plan as they wrote (Table 7.6) with the result that their mean for Generating Content rose in the second third of time while the C1 participants' mean fell. When C1 participants generated content later on they often seemed to be adding examples to support their main arguments compared to the B2s who were coming up with the main ideas (see Extract 7.6 a and b).

The C1 participants appeared to think specifically about the reader (see extracts 7.9) or about the suitability of language for the set task (see extracts 7.14 and Table 7.12) suggesting that they were more concerned with the impression that their text makes and aware of the impact that features such as choice of lexis has. They also thought about Linking Paragraphs and the impact that this has on clarity (extracts 7.13).

B2 participants tended to do less monitoring and revising of their texts, particularly towards the end of the production process (Tables 7.12, 7.13) while C1 participants tended to engage earlier with the monitoring processes. When it came to revision, the C1 participants seemed more willing to make changes to content rather than just to language errors and consideration was given the C1 participants to whether arguments made

sense and the impact of particular pieces of language on the reader (extracts 7.17 and 7.18).

In summary, there do seem to be differences between the amount that different processes are drawn on and from the verbal report data, differences in the aspects of writing that the B2 and C1 participants attended to within the processes.

Chapter Eight: Results from the Product Strand of the Study

8.1 Introduction

Chapter Seven set out the results from the process strand of the study. Chapter Eight will now set out the results from the product strand of the study and investigate the use of metadiscourse markers as indicators of discourse competence in the writing of candidates. Three of the research questions informed the product strand of the investigation. These are:

3. Which metadiscourse markers are used by candidates of the ESB ESOL International Examinations at levels B2 and C1 of the CEFR?
4. What are the functions of the metadiscourse markers used by candidates at level B2 and C1?
5. To what extent do the metadiscourse functions employed by candidates at levels B2 and C1 match the predictions made by the CEFR regarding the development of discourse competence in learners at these levels?

In order to explore these questions a number of hypotheses were generated which are set out in Table 6.8.

Section 8.2 sets out the descriptive statistics from the product strand of the study. Sections 8.3 to 8.10 will then address each hypothesis in turn. The results will be discussed in Chapter Nine in sections 9.3 to 9.5.

8.2 Descriptive Statistics

8.2.1 Recap on the text-selection process

As set out in Chapter Six, all of the candidate scripts were rated by a panel of individuals with experience of working with the CEFR and the results were analysed using FACETS (Linacre, 2008). Following the process set out in 6.5.8, a total of thirty scripts at C1 and thirty scripts at B2 were

selected for analysis. The scripts were then examined using the ‘find’ function in Word (Microsoft, 2013) to identify metadiscourse markers based on Hyland’s (2005) categories and lists.

Hylands’s 2005 scheme (see 4.6) was used as the basis for the analysis. Tables 8.1 and 8.2 show the different categories of metadiscourse markers in the scheme, the number of lexical items included initially and the items which were added during analysis of scripts. It is against this combined list of 495 items that the results for this strand were calculated.

Table 8.1 *Interactive Metadiscourse categories based on Hyland 2005 (2005, p. 218-224) and additional items.*

Type of marker	Metadiscourse Category (Hyland, 2005)	No. of lexical items in Hyland’s 2005 scheme	Example lexical items	Added items	Added items
Interactive metadiscourse	Code gloss	26	Called, I mean, i.e.	0	N/A
	Endophoric Markers	13	In this part	2	In the last paragraph, in the introduction
	Evidentials	7	Cited, quoted	0	N/A
	Sequencing	21	First, lastly	2	First and foremost, last but not least
	Label stages	19	To start with, so far	0	N/A
	Announce goals	14	Aim, goal, intend to	0	N/A
	Shift topic	14	Now, so, well	1	As regards to
	Transition markers	51	And, but, however	6	This means, which means, on the one hand, on the other side of the coin, every coin has two sides, in order to

Table 8.2 *Interactional Metadiscourse categories based on Hyland 2005 (2005, p. 218-224) and additional items.*

Type of marker	Metadiscourse Category (Hyland, 2005)	No. of lexical items in Hyland's 2005 scheme	Example lexical items	No. of additional items	Added items
Interactional metadiscourse	Attitude markers	64	Agree, essential	0	N/A
	Boosters	64	Always, definitely	0	N/A
	Self-mention	11	I, me, we	0	N/A
	Engagement markers	80	By the way, refer	0	N/A
	Hedges	99	About, fairly, tend	2	In most, personally
Total		483		12	
Overall total				495	

8.2.2 Types, tokens and means

Table 8.3 provides the overview of the results of the analysis in terms of descriptive statistics and provides a summary of the data (Dornyei, 2007, p. 209). One of the interesting findings here is that despite the fact that B2 participants were only required to produce 180-200 words versus the 250-280 words required in the C1 task, the mean length of response for the B2 candidates actually fell within the C1 task word requirement. This had the impact of making the two groups of writers close in terms of the number of words produced for each corpus. However, the fact that the C1 candidates wrote more may have distorted the results in that the higher-level candidates may have used more metadiscourse markers simply by virtue of having produced more words in total. Table 8.4 and Figures 8.1 and 8.2 attempt to correct for this effect by comparing the proportion of discourse as a percentage of the total text in each group.

Table 8.3. *Descriptive statistics for types and tokens.*

CEFR Level		Types	Tokens
B2	Mean	124.90	281.100
	Std. Deviation	22.179	53.21
	Median	124.50	277.00
	Minimum	82.00	169.00
	Maximum	176.00	386.00
C1	Mean	133.80	302.40
	Std. Deviation	23.33	53.99
	Median	132.50	297.50
	Minimum	97.00	225.00
	Maximum	192.00	452.00

Table 8.4. *Mean of metadiscourse markers as percentage of texts.*

	Mean length of Texts	Mean number of metadiscourse markers	Mean of metadiscourse markers as percentage of text
B2	281.1	33.33	4.22
C1	302.4	36.7	4.04

Figure 8.1. *Comparison of metadiscourse use by mean.*

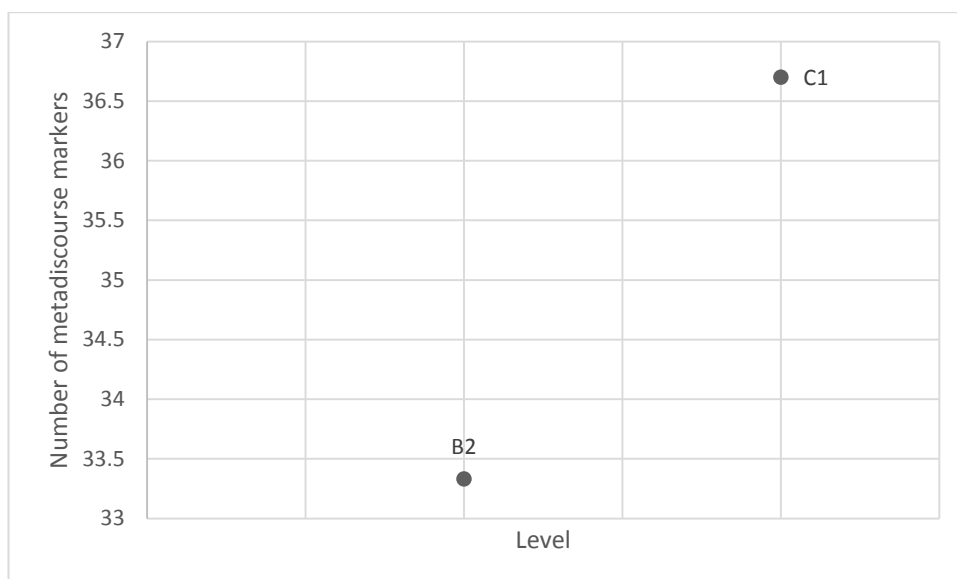


Figure 8.2. *Comparison of metadiscourse markers used as percentage of texts.*

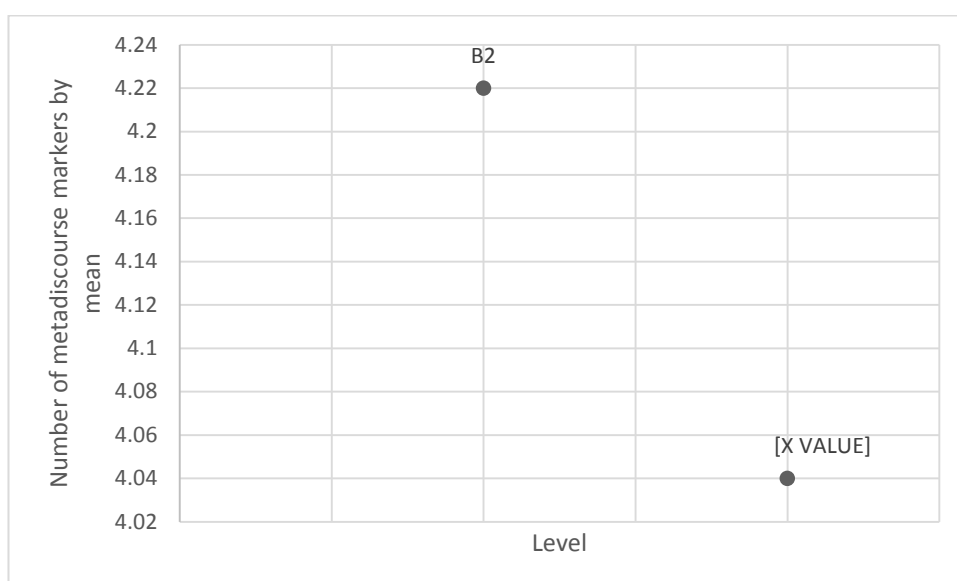


Table 8.4 and Figures 8.1 and 8.2 illustrate that even though the C1 participants used more metadiscourse markers, in terms of mean the number of metadiscourse markers used by C1 participants was lower in comparison to the B2s. Table 8.5 compares the descriptive statistics for the two types of metadiscourse marker identified by Hyland (2005),

interactive and interactional markers. See 6.1.1 for a description of this division.

Table 8.5. *Comparison of interactional and interactive markers as means and percentage of text.*

Level	Tokens	Interactive markers (Mean)	Interactional markers (Mean)	Interactive markers (%age)	Interactional markers (%age)
B2	281.1	643	396	2.29	1.41
C1	302.4	682	466	2.26	1.54

Table 8.5, similar to Table 8.4, shows that although the C1 candidates used more metadiscourse in their writing overall, when this was calculated as a percentage of text there was a different pattern. The use of interactive metadiscourse at C1 was lower in comparison to the use by B2 candidates. At the same time interactional metadiscourse used by the C1 candidates was higher by 0.71%. The full descriptive statistics for metadiscourse by category is in Appendix Eleven. However, due to the potential distortive effect of the length of the C1 texts the following sections (8.3 to 8.6) will review the results for each of the hypothesis but will draw on inferential statistics in order to identify features of statistical significance in the data.

8.3 Inferential statistics

Before statistical tests could be applied to the data it was necessary to determine whether or not the data collected from the B2 and C1 corpora was distributed normally. As identified by Connolly (2007, p. 206) and Dornyei (2007, p. 208) normal distribution is a prerequisite for parametric tests and where data is not distributed normally, non-parametric tests must be used.

In order to determine whether the data was normally distributed a one-sample Kolmogorov-Smirnov test was used. This test is recommended by Connolly (2007, p. 201-203) in order to establish whether an independent t-test can be used to analyse the data. Unlike the data from the process strand of the study, the distribution of the metadiscourse markers used in each category was found not to be normally distributed (see Table 8.6 below).

With the data found not to be distributed normally the Mann-Whitney U test which is a non-parametric alternative to an independent samples t-test was used for the reasons stated in 6.6.3.

Table 8.6. *Significance from one-sample Kolmogorov-Smirnov tests for metadiscourse markers by category.*

Category	Asymptotic Sig. (2-sided test)*
Code gloss	0.00
Endophorics	0.00
Evidentials	0.00
Sequencing	0.00
Label Stages	0.00
Announce Goals	0.00
Shift Topic	0.00
Transition	0.01
markers	
Attitude Markers	0.00
Boosters	0.00
Self-Mention	0.00
Engagement	0.00
Markers	
Hedges	0.00

* The significance level is .05

The following three sections review each of the research questions in turn and consider the results of the hypothesis related to each of the questions (see Table 6.8 for a summary).

8.4 Research Question Three

8.4.1 Hypothesis One

There is a significant difference between the total number of metadiscourse markers used by candidates at B2 and C1

The descriptive data suggested that there are differences in the total number of metadiscourse markers used by candidates at B2 and C1. Table 8.4 shows how the actual number of metadiscourse markers used by C1 candidates was lower than that of the B2 candidates when considered as a percentage of the overall text.

The Mann-Whitney U test was applied to the overall data to determine whether there was any statistically significant difference in the total number of metadiscourse markers used by B2 and C1 candidates. The medians for B2 and C1 were 27.2 and 33.2 respectively. The two groups did not differ significantly (Mann-Whitney $U = 351.0$, $B2 = C1 = 30$, $p < 0.05$ two-tailed). Therefore, the data rejects hypothesis one, which is that there is no statistical difference in the total number of metadiscourse markers used at B2 and C1, as was expected based on previous studies (see 4.7).

8.5 Research Question Four

8.5.1 Hypothesis Two

There is a difference in the proportion of interactive metadiscourse markers used between levels B2 and C1.

As with hypothesis one, the descriptive statistics suggested that there was a difference in the proportions of interactive metadiscourse markers used by the B2 and the C1 candidates. As shown in Table 8.7, the B2 candidates used a higher number of interactive markers in comparison to

the average length of their texts. The Mann-Whitney U test was applied to the data to determine whether the results were statistically significant. The results, shown in Table 8.7 below indicated no statistical significance.

Table 8.7. *Mann-Whitney U test on overall totals for interactive metadiscourse markers compared by level (B2 and C1).*

	Total
Chi-square	0.27
df	1
Asymp. Sig.	0.55

This result suggests that hypothesis two must be rejected as no statistical significance can be shown.

8.5.2 Hypothesis Three

There is a difference in the proportion of interactional metadiscourse markers used between levels B2 and C1.

As before, the descriptive statistics suggested that there was a difference in the proportions of interactional metadiscourse markers used by the B2 and the C1 candidates. As shown in Table 8.5, the B2 candidates used a higher number of interactional markers in comparison to their average text length. The Mann-Whitney U test was applied to the data to determine whether the differences were statistically significant. The results, shown in Table 8.8 below indicated no significance.

Table 8.8. *Mann-Whitney U test on overall totals for interactional metadiscourse markers compared by level (B2 and C1).*

	Total
Chi-square	0.27
df	1
Asymp. Sig.	0.56

As with hypothesis two, this result suggests that hypothesis three must be rejected as no statistical significance can be demonstrated in the differences between B2 and C1 uses of interactional metadiscourse markers in total.

8.5.3 Hypothesis Four: *There are differences in the way individual categories of metadiscourse marker are used between B2 and C1.*

As set out in 4.7, although there may be little in the way of difference in the overall numbers of metadiscourse markers in scripts at the two levels, studies show that there are variations within the individual categories (such as Transition Markers, Code Glosses etc.). Mann-Whitney U was used to test for the statistical significance of the observed differences. Table 8.9 shows the results for each interactive marker category while Table 8.10 contains the results for each interactional category. The Bonferroni Correction was applied in order to reduce the risk of Type 1 errors (Connolly, 2007, p. 197).

Table 8.9. *Results of Mann-Whitney U test for statistical significance of interactive metadiscourse markers by category.*

	Code Gloss	Endophorics	Evidentials	Sequencing	Label Stages	Announce Goals	Topic Shift	Transition Markers
U	461.00	450.00	465.00	424.00	495.00	434.50	420.00	514.50
Test Stat	0.17	0.00	0.59	-.40	0.77	-.61	-1.43	0.51
df	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Asymp . Sig.	0.867	1.000	0.557	0.689	0.442	0.544	0.154	0.339
r	0.02	0.00	0.08	0.05	0.10	0.08	0.18	0.07

p < 0.006 Bonferoni correction

r = Eta squared

Table 8.10. *Mann-Whitney U test for statistical significance of interactional metadiscourse markers by category.*

	Attitude Markers	Boosters	Self Mention	Engagement Markers	Hedges
U	473.50	581.00	302.00	511.50	581.00
Test Stat	0.37	2.02	-2.21	0.95	1.97
df	1.00	1.00	1.00	1.00	1.00
Asymp. Sig.	0.713	0.044	0.027	0.342	0.049
r	0.05	0.26	0.29	0.12	0.25

p < 0.01 Bonferroni correction

r = Eta squared

The results show that there were no statistically significant differences in the interactive categories or interactional categories after the Bonferroni Correction was applied. Use of the correction can result in type two errors, where differences are not detected (Pallant, 2005, p. 200). Three interactional metadiscourse categories demonstrated results which were significant at the 0.05 level: Boosters, Self-mention and Hedges. All three also demonstrated strong relationships through their effect scores. This would support the results from the descriptive statistics (see Table 8.4) which showed an increase in the proportion of interactional metadiscourse markers used by C1 candidates. However, the number of tests used necessitates a more stringent level of significance so the hypothesis must be rejected.

In order to explore hypothesis four further, the range of different markers used in each category was explored. Each different individual metadiscourse marker used at each level was counted and Table 8.11 shows the results. These show that in terms of the interactive markers, the B2 candidates used the same or more markers in six categories. The C1 candidates only used more types of Code Glosses and Transition Markers. However, in the interactional categories, the C1 candidates made use of more types of metadiscourse markers in three of the five categories.

Table 8.11. *Descriptive statistics for the number of different metadiscourse marker by category.*

Category		Number of exponents used	
		B2	C1
Interactive metadiscourse	Code gloss	8	12
	Endophoric markers	2	2
	Evidentials	1	1
	Sequencing	11	9
	Label stages	5	4
	Announce goals	2	1
	Shift topic	2	0
	Transition markers	25	32
	Attitude Markers	9	10
Interactional metadiscourse	Boosters	16	22
	Self-Mention	6	5
	Engagement Markers	8	8
	Hedges	19	27

The Mann-Whitney U test for statistical significance was applied to the range of individual types of metadiscourse markers used in the categories of Boosters, Self-mention and Hedges in order to determine whether there was a difference in the range of types used. Self-mention and Hedges showed differences as shown in Table 8.12 below when a more lenient test of significance was used ($p < 0.05$). When the Bonferroni Correction was applied, no difference in terms of the range of types used was found.

Table 8.12. *Mann-Whitney U test for range of markers used for Boosters Self-mention and Hedges.*

	Boosters	Self-Mention	Hedges
U	574.00	309.50	585.00
Test Stat	1.94	-2.16	2.05
df	1	1	1
Asymp. Sig.	0.05	0.03	0.04
Effect Size ^a	0.25	0.28	0.26
p < 0.02 Bonferroni correction			

^a Eta squared

Based on the results, hypothesis four is rejected as there are no statistically significant differences in the way the B2 and C1 candidates use metadiscourse markers by individual category. There is some difference in the way C1 candidates use Boosters, Self-mention and Hedges and in the range of linguistic exponents used but no relationship can be claimed.

8.6 Research Question Five

8.6.1 Hypothesis Five

There are differences in the way individual types of metadiscourse markers are used by B2 and C1 candidates.

As discussed in 4.7, the CEFR predicts that there will be differences in the ways in which learners at the B2 and C1 levels will use metadiscourse markers. These are:

- At the B2/B2+ level the “use a limited number of cohesive devices to link sentences together smoothly into clear, connected discourse; use a variety of linking words efficiently to mark clearly the relationships between ideas; develop an argument systematically

with appropriate highlighting of significant points, and relevant supporting detail.” (Council of Europe, 2001, p. 35)

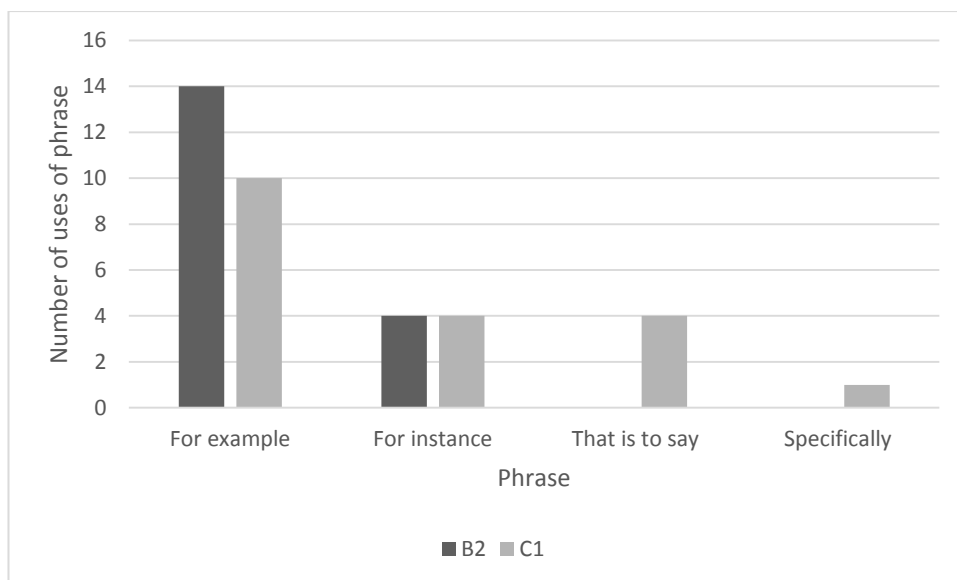
- At the C1 level these discourse skills will “continue to be evident” and learners will “show controlled use of organisational patterns, connectors and cohesive devices”. (Council of Europe, 2001, p. 36)

For hypothesis five the individual lexical exponents of metadiscourse markers were examined as set out in 6.6.3. A number of categories had such small numbers of markers used that they were not considered in detail: Endophorics, Evidentials, Label Stages, Announce Goals, Shift Topic and Attitude Markers. In the following sections, each of the remaining categories are considered in turn.

8.6.1.1 Code Glosses (interactive)

As shown in Table 8.11, the C1 candidates used more linguistic exponents to carry out the Code Gloss function. Figure 8.3 shows that the most popular exponent was ‘for example’. This piece of language is identified by the English Vocabulary Profile (Cambridge English Language Assessment et al, 2014) as being a piece of language from the A1 level, and could be therefore considered a good example of a ‘lexical teddy bear’ (Carlsen, 2010. See 4.7). It is therefore interesting to note the slight decline in use by C1 candidates. ‘That is to say’ and ‘specifically’, which the English Vocabulary Profile (Cambridge English Language Assessment et al, 2014) identifies as being C1 in level are not used by the B2 candidates at all. Taken with the greater range of exponents used by the C1 candidates, as shown in Table 8.11, this data seems to point to changes in the way higher-level candidates carry out the function of Code Glossing.

Figure 8.3. *Comparison of 4 types of Code Gloss*



8.6.1.2 Sequencing (interactive)

As shown in Table 8.13, the B2 candidates used a wider range of Sequencing metadiscourse markers and also used more (44 examples of Sequencing compared to 38 used by the C1 candidates).

Table 8.13. *Comparison of most frequently used Sequencing items*

	Numbers of use by B2 Candidates	Numbers of use by C1 Candidates
First of all	13	17
Firstly	6	1
To begin	4	6
Finally	8	3
Last but not least	6	5

The use of ‘last but not least’ is described by the English Vocabulary Profile (Cambridge English Language Assessment et al, 2014) as “something that you say to introduce the last person or thing on a list”. In the essays at B2 and C1 in the study the phrase was often used to introduce the final argument in a paragraph and used for Sequencing. The

two differences were examples (viii) and (x), both in C1 scripts where the phrase was used in the conclusion and these instances appear to be more correct in their usage than the other instances. Example (vii) is also interesting because it shows a departure from the set phrase and a more complex combination of phrases.

Figure 8.4. *Concordance lines for 'last but not least'*

B2

- i. miss your parents and your friends. LAST BUT NOT LEAST is that you feel lonely because you are new
- ii. or basketball in their schools. LAST BUT NOT LEAST some people think that with sports at schools children
- iii. not have enough money to do this. LAST BUT NOT LEAST there always exist the danger of your own health
- iv. you think differently, as a older. LAST BUT NOT LEAST you can have a option of the world. I mean that you live
- v. own homes and they do not live under poverty. LAST BUT NOT LEAST, in the past a lot of wars happened between
- vi. because of the financial crisis. LAST BUT NOT LEAST, parents can't advice their children with a view to deal

C1

- vii. it is said you feel foreigner between foreigners. LAST but as far as I am concerned NOT LEAST is that you have to live alone
- viii. but unfortunately there is a lack of them. LAST BUT NOT LEAST, I think that it is very good for everyone to live for m
- ix. they will not achieve to live. LAST BUT NOT LEAST, it is of utmost significance to point out that older
- x. do we eat and how often do we exercise. LAST BUT NOT LEAST, no matter how much we want longevity we
- xi. memorable experiences. LAST BUT NOT LEAST, you may have to opportunity to find easier a job there

The greater and wider use of Sequencing markers by B2 candidates could indicate a more additive, knowledge-telling (Bereiter & Scardamalia, 1987) approach to writing within which the candidates are forced to keep linking their text as they write while the C1 participants who have already generated much of their content and organised the text could rely on few, more planned Sequence Markers as suggested by the process strand (see 7.5.2 and 7.5.4).

8.6.1.3 Transition Markers (interactive)

The Transition Marker category had by far the most instances of use by all candidates which is unsurprising given the inclusion of very high frequency words such as ‘and’, ‘but’ and ‘so’ in the category. The category demonstrated no statistical difference in terms of overall metadiscourse markers (as shown in Table 8.7) and as reported in 8.5.3, there was no statistical difference in terms of the range of types of Transition Markers used in the category although the C1 candidates were using a larger range of markers (see Table 8.11).

The Transitional Markers were considered according to their means and the results are in Table 8.14. The majority of the items used by B2 learners are rated by the English Vocabulary Profile (Cambridge English Language Assessment et al, 2014) as being A1 – A2 in level while the items in the C1 list are rated as largely B1 – B2. What this may suggest is that while the B2 learners are likely to be aware of the range of items they *could* use for the Transition Marker functions, they often rely on a core of very high frequency, simple items, in other words the “lexical teddy bears” (Hasselgren, 1994).

Table 8.14. *Use of different Transition Markers according to mean.*

B2 > C1	B2 < C1
also	Besides*
although	consequently*
because	even though*
but	furthermore
so	however
	in addition
	in order to
	moreover
	nevertheless*
	therefore*
	thus*
	whereas*

*markers not used by any B2 candidates in their scripts.

Examination of the use of some of the markers also demonstrate some interesting patterns. Figure 8.5 shows that both instances of 'although' by the B2 candidates contained errors. Example (i) is functionally correct but wrong in terms of punctuation while (ii) is not functionally correct. The C1 candidates' examples are functionally correct but also demonstrate more sophistication in their use, for example the use of 'however' in lines (viii) and (ix) with 'however' used to signal the shift in the argument and 'although' used to recap the previous point before introducing the differing proposition.

Figure 8.5. *Concordance lines for 'although' from B2 and C1 candidates.*

B2

- i. This will benefit the government at the economic side. **ALTHOUGH**, a longer life has disadvantages, too. It costs the government
- ii. **ALTHOUGH**, medicines now harm less the human body and they can be effective in a very little time.

C1

- iii. far outweigh disadvantages. **ALTHOUGH** it is claimed that travelling overseas in order to study is not so
- iv. On the other hand, there are some negative impacts too, **ALTHOUGH** less important. For example, overpopulation is a negative consequence
- v. Today's society is faced with a fact that, **ALTHOUGH** sounds very positive and hopeful, it forces everyone to think why
- vi. so from all points of view it is a real enjoyment. **ALTHOUGH** the longer life expectancy may have negative aspects for the society
- vii. discover all the beauty that this planet has to offer. **ALTHOUGH** there are both negative and positive impacts in the increase in life
- viii. the process they have been doing. However, **ALTHOUGH** these issues are proved to be a constant and rising danger for today
- ix. more years than they did in the very past. However, **ALTHOUGH** this fact seems to be marvelous and mysterious, it can be explained
- x. how do they live there, which broaden your horizons. **ALTHOUGH**, there are others who think that studying abroad is not a

Finally, in the case of 'but' 38% of the uses by B2 candidates are to start sentences, compared to 15% by the C1 candidates suggesting that there is a difference between the way candidates at these levels use the item.

8.6.1.4 Boosters (*interactional*)

As shown in 8.5.3 above, the Mann-Whitney U test revealed no significant difference between B2 and C1 candidate scripts in the use of Boosters. Table 8.4 shows that the C1 candidates' mean for Boosters was much higher than that of the B2 candidates (2.10 and 1.27 respectively) while the C1 candidates used 22 different lexical exponents for the function of Boosting compared to 16 exponents used by the B2 candidates. The higher number and range of Boosters at the higher level of proficiency could be indicative of more confidence about the propositions being made

but also illustrate more willingness to go beyond the tried-and-tested exponents employed by candidates at the lower level.

Table 8.15. *Examples of Boosters used by B2 and C1 candidates with English Profile level of exponents (in brackets).*

B2 Candidates	Used by B2 & C1 candidates	C1 Candidates
certainly (A2) prove (B1) truly (C1)	always (A1/A2) believe (A2) of course (B1) really (A2)	evident (B2) no doubt (C1) obviously (B1) surely (B2) undeniable (C1) without a doubt (B2)

Table 8.15 illustrates how the C1 candidates appear to be stretching their language beyond the lower-level exponents in order to carry out the function. As was discussed in 8.1.6.3, what is of interest is that the B2 candidates may be familiar with many of these lexical items, as the English Vocabulary Profile (Cambridge English Language Assessment et al, 2014) ranking would suggest but that they have opted to use a narrower and simpler set of exponents to carry out the function.

Table 8.16. *Words occurring to the left and right of Booster ‘of course’ in B2 and C1 candidate scripts.*

B2 Candidates		
FULL STOP (4)	of course	
and (1)		
		this (1)
C1 Candidates		
FULL STOP (3)	of course	this (3)
and (4)		it (1)
This, (1)		they (1)

There can also be found differences in the way the lexis is used. Table 8.16 gives an example of the use of ‘of course’ by B2 and C1 candidates.

80% of the use the phrase by B2 candidates was at the start of a sentence to introduce a proposition compared to 27% of C1 candidates. However, 76% of C1 candidates used 'of course' with a determiner or pronoun as a form of substitution, suggesting that the phrase was being used to comment on an observation made earlier in the text.

8.6.1.5 Self-mention (interactional)

Most of the exponents in the Self-mention category had high levels of use. Table 8.17 shows that the B2 candidates tended to use more of each lexical exponent but when these were tested using the Mann-Whitney U test none of these differences were found to be statistically significant once the Bonferroni Correction had been applied.

Table 8.17. *Mann-Whitney U test results by individual lexical exponents for Self-mention.*

Item	Level	Counts	Mean	SD	Test Statistic	df	Sig*
I	B2	38.00	1.27	1.66	0.00	1.00	0.99
	C1	33.00	1.10	1.42			
we	B2	54.00	1.80	2.41	5.29	1.00	0.02
	C1	32.00	1.07	3.27			
my	B2	8.00	0.27	0.45	2.74	1.00	0.98
	C1	3.00	0.10	0.31			
our	B2	40.00	1.33	1.97	0.01	1.00	0.94
	C1	36.00	1.20	1.81			
us	B2	17.00	0.57	1.10	4.25	1.00	0.04
	C1	4.00	0.13	0.57			

$p > 0.01$ Bonferroni Correction

Despite the lack of statistical difference, it can be seen that there was a decline in the use of these markers in essay writing by the C1 candidates.

This might be indicative of a developing sense of genre awareness and of the need to avoid such pronouns in an essay (see 4.6.2.3).

8.6.1.6 .Hedges (interactional)

As shown in Table 8.11, the category of Hedges, showed that C1 candidates used a wider range than the B2 candidates. There were also observable (non-significant) differences in the category, suggesting that there may be differences to be found in the way B2 and C1 candidates carry out this function were a larger sample to be examined. Comparison of means between the B2 and C1 candidates use of Hedges is shown in Table 8.18 below.

Table 8.18 *Comparison of Hedges used by B2 and C1 candidates based on mean.*

B2 > C1	C1 > B2
almost*	claim
often	generally*
in my opinion	may
	might
	mainly*
	personally*
	tend to

* indicates that there are no counts of this word at the other level.

The means demonstrate that the C1 candidates employed a wider range of markers for Hedges but it also hints at the tendency for B2-level candidates to make use of 'simpler' items. 'Almost' is rated by the English Vocabulary Profile (Cambridge English Language Assessment et al, 2014) as being A2 in level and while the use of 'often' in the meaning of being 'normal or true' is rated at B2, the word itself is very high frequency and learners may have been aware of it since the A1 level of the CEFR with the meaning of 'regularly'. The words used by the C1 candidates, while

often ranking as B1 or B2 according to the English Vocabulary Profile may not have been acquired as early or be quite as high in terms of frequency. It is possible that B2 candidates stick to familiar language to carry out these functions even when they know of alternatives.

8.7 Conclusions

In this chapter we have established the following based on the analysis of the statistical information from the product strand of the study:

- Hypothesis one is rejected as expected. There is no statistically significant difference between the total number of metadiscourse markers used by candidates at levels B2 and C1
- Hypothesis two is rejected. There is no statistically significant difference in the proportion of interactive metadiscourse markers used by candidates between levels B2 and C1.
- Hypothesis three is rejected. There is no statistically significant difference in the proportion of interactional metadiscourse markers used by candidates between levels B2 and C1
- Hypothesis four is rejected. There do not appear to be differences in the way individual categories of metadiscourse markers are used by candidates between levels B2 and C1. However, when the test for significance is relaxed interactional categories (Boosters, Self-mention and Hedges) show differences. All three categories also demonstrate large effect size.
- For hypothesis five there appear to be differences in the way individual lexical exponents are used by B2 and C1 candidates. This appears to suggest that Hasselgren (1994) and Carlsen's (2010) observation that lower-level learners tend to rely on a set of familiar expressions is accurate. The way in which some expressions **are used** (e.g. although, but, last but not least) also seems to change between the B2 and C1 scripts.

These results and the hypotheses will be discussed in relation to the research questions in Chapter Nine.

Chapter Nine: Discussion

9.1 Introduction

This chapter will explore and draw together the results set out in Chapters Seven and Eight and will attempt to answer the research questions based on this data. The chapter is structured using the research questions (sections 9.2 – 9.5) but will draw on both strands. Chapter Ten will draw conclusions on the aims of the study and discuss the limitations of the study and potential areas for further research.

An important point to make before beginning the discussion, and one which will be reiterated when limitations of the study are discussed (see 10.3) is the issue of the multidimensionality of the CEFR and of skills. De Jong (2009) suggests that quality and quantity of performance may not be consistent and that the CEFR itself points this out (Council of Europe, 2001, p. 169) when it states that:

“all knowledge of a language is partial, however much of a ‘mother tongue’ or ‘native language’ it seems to be. (...) In addition, a given individual never has equal mastery of the different component parts of the language in question, for example, of oral and written skills, or of comprehension and interpretation compared to production skills.”

Performances should always be considered from this multi-dimensional perspective. The participants in the qualitative strand of the study have all passed qualifications appropriate to the level which they represent (and qualifications from awarding organisations independent of ESB). Likewise, the scripts chosen for the quantitative strand of the study are all from passing candidates and have been rated independently but both the process and the product sets of data should be seen as being on a

multidimensional axis and some traits may be less-developed than others while some areas may seem to have more in common with a higher level of proficiency. This is particularly the case when one considers the overlap between the higher end of B2 level and the C1 level. As explored in 3.7, the CEFR does not provide much guidance as to how discourse competence develops between the two levels (e.g. at C1 “the discourse skills characterising the previous band continue to be evident” (Council of Europe, 2001, p. 36). These issues will be returned to in the limitations and suggestions for further research.

9.2 Research Questions One and Two:

What cognitive phases do candidates at levels B2 and C1 appear to employ when composing timed essay tasks?

To what extent does the timed essay question format used in the ESB ESOL International Examinations elicit the cognitive processing that models predict at levels B2 and C1?

Table 7.1 shows that all of the categories identified in the coding process were used by the participants at B2 and C1. Table 7.4 combined these categories into Field’s phases (2004, p. 329-339). The results show that the participants at B2 and C1 drew on all of the categories and phases to greater or lesser degrees. This is not particularly surprising. Shaw and Weir (2007, p. 49-62) discuss how the cognitive requirements of writing change across the levels of the Cambridge ESOL Examinations but it is not the presence of a particular cognitive phase that distinguishes a performance at B2 or C1 but *how* it is carried out. For example, in organisation at B2, Shaw and Weir state that “students are advised...to make a plan for their answer, noting what to include in each paragraph” while at the C1 level “strong candidates organised and structured their report well...weaker candidates failed to plan their answers and often tried to include every piece of information.” (2007, p. 55-56). If we accept the

results which show all of the phases being used by the B2 and C1 participants then the issue becomes a question of not *what* phases are being used but *how* the phases are being used at the different levels.

The next sections will explore each of the categories of analysis and examine this question of how the participants carried out their processing. Each category will be discussed, first in terms of Research Question One then in terms of Research Question Two.

9.2.1.1 Macroplanning (Task Assessment, Generating Content, Consider Audience and Word Count)

Field's Macroplanning phase (2004, p. 329) showed no statistical difference between the B2 and C1 participants either as an overall phase (Table 7.5) or in the subcategories (Table 7.3) which make up the phase. However, when the phases were analysed according to whether they occurred at the beginning, the middle or end of the period available for the writing task, the Macroplanning phase did show a difference in the final third of writing time when participants were working towards concluding their essays (see section 7.4). Investigation by category did not reveal statistically significant differences between the two groups. In terms of means, Table 7.6 shows that B2 participants were Generating Content more in this time section than the C1 participants but this was not statistically significant. Figure 7.2 drawn from the post-interview question about how participants prepared to write their compositions suggests together with the data from Table 7.5 that the C1 participants devoted more time to preparation in the earlier stages of writing including the generating of ideas and consideration of what language they might need. The comments such as in extracts 7.5 made by the C1 participants support this interpretation while the comments from the B2 participants suggest a much more 'hand to mouth' style of production by the lower-level participants. Extract 9.1 is also an example of this.

Extract 9.1

I was erm trying to <\$=> to think some erm <\$=> some reasons oh no <\$=> some positive impacts but I don't erm <\$=> I have not think <\$=> have not thought a lot of <\$E> pause 2.5 seconds <\\$E> er but I have remembered that er people who live at <\$=> lived years ago er they don't live a lot they <\$E> pause <\\$E> erm live er in fifty years <\$E> pause <\\$E> like that.

(Participant 5, B2 <16:03>)

It also appears from the verbal report data such as the extracts in 7.2 that when C1 participants were generating ideas beyond the first third of writing time they were often seeking examples to support the main points rather than coming up with main ideas (as in the case of the B2 participant in extract 9.1). Another point which connects to Table 7.7 and the instances of note-taking prior to writing was that when making preparatory notes, C1 participants were more likely to write down their own ideas and keywords rather than just words from the question.

In terms of thinking about the audience and the demands of the task, Figure 7.4 suggests that while all of the participants had the notion that there was a formal audience for the piece of writing the C1 participants had a clearer notion of for whom they were writing and more often identified this with an assessor or a teacher (i.e. someone who would make a judgement about their writing). This data from the post-interview is supported by the comments made by C1 writers throughout the process such as in extracts 7.9 but also by comments made by C1 participants in the Searching for Lexis category (extracts 7.14). This suggests that consideration of the audience is an iterative process and one which is connected to different aspects of text composition. It also suggests that the impact of the text on the reader is a consideration that C1 participants are more likely to take into account when composing.

9.2.1.2 *Macroplanning and the models*

Based on the above it can be seen that C1 participants take more note of the target readership. Field (2004, p. 329), as discussed in 2.7.1, suggests that more proficient writers will do this and consider the style of writing that a particular audience expects. This finding would also support Hyland's (2002, p. 26) observation that second language writers pay less attention to the goals of their texts while suggesting that this does change as learners progress in proficiency. This also fits with the observation in 7.5.7 that the C1 participants were more concerned with finding the 'right word' in that the clearer concept of audience provided criteria for this decision.

Field (2004, p. 329), following Hayes and Flower (1980) observed that the cognitive phases were not linear and the findings indicate recursive and simultaneous phases. For example, categories for macroplanning were used in each third of writing time. Furthermore, notions of readership and impact (as discussed above) were used by C1 participants in particular to inform decisions about aspects such as choice of language.

Another interesting feature from the data comes from Figure 7.2 where B2 participants reported that their pre-planning involved comparing the set task with previous essays they had produced. Extracts 7.3 and 7.4 record B2 participants saying that they were 'familiar' with the essay. However, later on the participants reported trying to come up with ideas:

Extracts 9.2

(a)

<\$2> I'm trying to find some positive and negative impacts of this increase of longevity <\$E> pause <\\$E> and so I'm trying to think because <\$E> pause <\\$E> I haven't thought of it before. <\$E> laugh <\\$E> so this is my first time thinking about it. And I'm trying to find some impacts positive and negative.

(Participant 4, B2 <11:10>)

(b)

<\$2> Okay. I finished with the first topic. I wanted to write about the medicine and now I just started thinking of the next thing that helps our <\$=> us to have a longer life expectancy. <\$E> pause <\\$E> I haven't thought of it yet.

(Participant 7, B2 <10:02>)

If we conclude from these extracts that the topic was unfamiliar then perhaps this suggests that what the participants were expressing familiarity with was the type of essay and the 'advantage/disadvantage' question. This familiarity might assist in pre-planning the structure of the essay and in assigning topics to paragraphs but not in the generating of content, hence these participants were still generating content simultaneously as they wrote. Also problematic is the notion of familiarity of topic, in that this does not mean that a candidate will have all the ideas they need to write the essay automatically to hand.

The CEFR predicts that candidates at C1 will "produce clear, well-structured texts" (Council of Europe, 2001, p. 24) and the findings in Macroplanning suggest that the C1 participants did indeed think more about the overall purpose and structure of their texts, thereby presumably demonstrating the developing discourse proficiency predicted by the CEFR.

The way in which the B2 participants generated content links with Bereiter and Scardamalia's (1987) models of knowledge telling and knowledge transforming. The models predict that less competent writers will be more concerned with questions around *what* they will write rather than *how* they will write it. Table 7.5 together with extracts 7.6 show how the B2 participants tended to be generating their ideas and then immediately writing them. An example of this can be seen in extract 9.3 below where a B2 participant towards the end of the second third of writing time was

forced to change the introduction to the paragraph because she had been unable to come up with a second advantage of longevity.

The tendency for B2 participants to knowledge tell (Bereiter & Scardamalia, 1987) may also explain some of the findings from the product strand of the study. If B2 writers are primarily concerned with generating ideas throughout the process, this would leave fewer cognitive resources in the working memory for producing text. The findings, such as those in 8.5 and 8.6, that B2 writers often used higher-frequency or lower level lexis (based on the English Vocabulary Profile (Cambridge English Language Assessment et al, 2014)) to carry out interactive and interactional functions, could be a result of the preoccupation with content, despite there being a range of lexical alternatives which a B2 writer could choose to use.

Extract 9.3

<\$2> <\$E> pause <\\$E> I couldn't find the second elaboration.

<\$1> Okay.

<\$2> So I rewrote the start of the second erm sentence and make it

<\$E> pause <\\$E>+

<\$1> one?

<\$2> +just one+

<\$1> Okay.

<\$2> +now I'm now moving on to the disadvantages.

(Participant 10, B2 <21:14>)

The behaviour of the B2 participants as illustrated by the examples above suggests that there is a greater propensity to simply write until out of ideas and then move on to the next task requirement. This would suggest that these writers are unlikely to flag ideas as being more or less prominent for the reader. The fact that the B2 writers tended to generate content as they went along also implies that these writers imposed greater cognitive

demands on themselves during the writing process as discussed in 2.7.2. The observation by Gathercole and Baddeley (1993, p. 5) that Working Memory has a limited capacity suggests that by forcing themselves to generate content at the same time as producing texts, weaker candidates will struggle due to the high demands of both activities. An anecdotal observation from the administration of the verbal reports emerged in the post-writing feedback to candidates where the researcher noticed that serious language errors and lapses tended to appear where learners stated that they had been struggling to generate content. This happened in the case of participant 6 (C1) who reported the following in the post-interview:

Extract 9.4

<\$1> Erm <\$=> Right you got to the top of here <\$E> pause <\\$E>
of page two and there was a very long pause
<\$2> Yes
<\$1> +about that point <\$E> pause <\\$E> a lot of pausing and a lot
of rubbing out
<\$2> Yes
<\$E> prompt at 55:27 <\\$E> <\$1> +then writing again <\$E> pause
<\\$E> could you remember what you were thinking about at that
point? <\$E> inaudible <\\$E>
<\$2> Erm I was brainstorming ideas but then I thought it was not
correct <\$=> incorrect because it was a bit out of the topic <\$E>
pause <\\$E>
<\$1> Yeah
<\$2> +that's why I was erasing.
<\$1> Okay erm <\$=> <\$E> pause <\\$E> and then around sixteen
to eighteen minutes you stuck around the same sort of space and
thinking about the task <\$E> pause <\\$E> and again there was
another long pause <\$E> pause <\\$E> what did you mean by that?

<\$2> Erm <\$=> about the positive and negative impacts because
erm <\$=> I=I couldn't come up with an idea <\$E> pause <\\$E> and
so I was <\$E> pause <\\$E> stuck but I was thinking about it.

<\$1> Okay erm <\$=> and then at the end of this very long pause
around eighteen minutes to twenty minutes you said "I thought of
them I organised them and now I'm writing them down" <\$E> pause
<\\$E> Okay.

<\$2> That's why I was stuck because <\$E> pause <\\$E> erm <\$=>
when I had found what I had write I=I write down then organise it in
the correct order.

(Post-Interview - Participant 6, C1 <55:27>)

Participant 6 was one of the C1 participants who started writing immediately (see field notes for participant 6 in Appendix Seven). In general, this participant displayed a good level of accuracy in his writing but at this point not only was there heavy erasing of text but a number of errors appeared in the writing. The observation here is that prior generation of content appears to assist the C1 participants in their writing and allows them to dedicate working memory to language and matters of audience while the B2 participants employ their cognitive resources to come up with the content as the knowledge telling model predicts.

As noted by Weir (2005a, p. 2) the CEFR provides very little explanation of what cognitive processing is required by candidates at different levels. It does state that by the C1 level learners should be able to "use language flexibly and effectively for social, academic and professional purposes" (Council of Europe, 2001, p. 36) and the findings of the study suggest that in terms of microplanning, the C1 participants were able to give more attention to the intended audience and had a clearer notion of for whom they were writing.

9.2.2.1 Organisation

Organisation did not return any statistically significant findings but the data did show observable differences. Table 7.6 shows how C1 participants did more organisation in the first third of writing time while the extracts in 7.11 highlight how participants thought about the overall structure of their writing. All of the participants used the title of the essay to structure what they wrote, effectively dividing their work into a section which explained why longevity had increased before going on to state the advantages and disadvantages and then producing a conclusion. This suggests that the essay structure had been taught to the learners along with basic structuring techniques and that it was well-learned. What can be seen from the discussion in the previous section on Generating Content, is that for the B2 participants structuring was a superficial process which involved dividing up the essay into parts rather than generating content and organising it. In contrast, participant 1 (Extract 9.5 a) did not appear to separate the writing process from planning, yet clearly considered that both had taken place. The same can be observed in the comments from participant 3 (extract 9.5 b).

Extract 9.5

(a)

<\$2> Er of course <\$E> pause <\\$E> first of all I read very well er all the task er I try to understand exactly what it wants <\$E> pause <\\$E> and then I try to er, <\$=> in my mind I try to separate all the paragraphs to see how many words will I use. Now I know exactly what to write in each paragraph for example. Er that's what I do when I <\$E> pause <\\$E> <\$=> when I write <\$E> pause <\\$E> and <\$E> pause <\$E> erm all my essays I start them the same way, the same style and then er it complicated in my mind I don't know <\$E> pause <\\$E> whatever I think I write. So, that my first paragraph and if I try to <\$E> pause <\\$E> think just as much as I

can without being boring <\$E> pause <\\$E> and always staying in the task.

(Participant 1, C1 <02:04>)

(b)

<\$2> Erm <\$E> pause <\\$E> I'm starting to write the prologue

<\$=> the er starting <\$=> prologue okay+

<\$1> Mmhmm.

<\$2> I'm doing er quite a wrong thing I'm <\$E> pause <\\$E> <\$=> I

start to think the erm <\$=> what er the general erm topic.

<\$1> Mmhmm.

<\$2> And I quite copy from the erm instructed <\$E> laughs <\\$E>

so this is what I've been doing up to here and I've meanwhile I've

been thinking of the ideas. I've been brainstorming on what I can

write.

(Participant 3, C1 <02:35>)

The B2 writers who did not plan did not report such concerns with structure or brainstorming and appeared to be more concerned with the immediate production of text. This seems to fit with Hyland's observation that second language writers often devote less time to planning than native speakers (2002, p. 26) but the fact that for some of the C1 participants the phases of Planning and Translation appear to merge supports the view Nystrand puts forward (Kellogg, 1999, p. 28) that it is difficult to distinguish between these phases and that the relationship between planning and translation may be co-dependent. Certainly those participants who were planning were also writing out their notes albeit in a form which was not intended for public reading. Another interesting note is that one B2 participant reported similar behaviour to many of the C1s in that she began writing but stopped and proceeded to plan her work (Participant 10).

9.2.2.2 Organisation and the models

From what has just been stated above it can be seen that Organisation incorporated decisions about the genre and the task for all the participants (Field, 2004, p. 329) but it was mainly the C1 participants who also considered the relationships of the content that they generated in this phase. This fits with the observations about Macroplanning, in that if B2 participants did not generate their content at the beginning then they could not organise it. The decline in Organisation as a phase after the first third of time can also be explained by the fact that once participants had made decisions about the structure of their texts these decisions became relatively fixed. Later references to Organisation were participant 2 who at <56:50> was beginning to worry about time and wondered whether a conclusion was necessary and participant 6 who started thinking about the conclusion early on in the writing process.

Kellogg (1999) notes that when writers are familiar with a particular genre they use pre-existing schema and thereby free up working memory to focus on other requirements of the task. This seems to fit with the behaviour of both groups of participants in the study who used the set question and their familiarity with the conventions of the advantage/disadvantage rhetorical structure to create the macro-organisation and to assist with paragraphing. However, because the B2 participants did not include Generating Content or the organisation of this content in their Macroplanning this may have resulted in little in the way of gain for these participants. This would fit with Johnson's argument (2012) that pre-task planning has little positive impact on lower-level learners. Any processing capacity in the working memory which B2 participants had freed up by having an overall structure was then taken up by generating content and composing text.

An argument put forward by Field is that "with experienced writers, it is sometimes execution that begins the whole writing process rather than

prior planning" (2004, p. 330) and it was observed during the pilot study that many of the native speakers started writing their compositions directly. Table 7.8 shows that of the B2 participants only two made plans while the other four began writing their essays immediately. This was also the case with three of the C1 participants, however the C1 writers often paused directly after writing part of the introduction to generate content and plan the essay or else appeared to have made the decisions while engaged in writing.

As previously discussed, it seems that the B2 participants often generated ideas and organised them as they wrote. This resulted in situations such as that illustrated by extract 9.3 above where text had to be re-written because the participant was unable to generate a second advantage. This would fit with the knowledge telling model (Bereiter & Scardamalia, 1987) while the C1 participants were more willing to think about how to convey their ideas in the text. The additive nature of the B2 participants' writing could also explain the more extensive use of interactive markers such as Sequencing markers identified in 8.6.1.2 of the product stand of the study. If the B2 writers were generating content as they went along then they may have had to constantly keep signalling the relationship of the new ideas whereas the C1 writers who had an overview of these relationships due to their pre-planning could use such forms more sparingly.

The CEFR, although poor at specifying the types of genres which learners are to produce at different levels does identify the essay as a particular text type and even includes essay writing as an illustration of competences (Council of Europe, 2001, p. 62). The CEFR suggests the advantage/disadvantage essay as an example of what B2 level learners are likely to be able to do and are often taught. As such learners are often expected to be familiar with this rhetorical structure.

The comments by the participants in the study and the observed planning described above suggests that the participants were indeed familiar with the form of an essay. All of them were aware that the language is more 'formal' than in letters and of how to structure the piece of work. However, there were some interesting points. Participant 9 (B2) produced what was clearly an essay but started the work with 'Dear Sir or Madam' and concluded it with 'Sincerely yours'. Participant 3 (C1) asked "Is this an article?". Two B2 and two C1 participants suggested that the essay was being written like a newspaper or a magazine article. This suggests that while the learners were aware of the form and conventions of the essay, probably from teaching, they were not always aware of the communicative purpose of such a text or of how it may differ from other similar texts. The findings in 8.6.5.1 (instances of Self-mention) indicate that C1 writers may be increasingly aware of some genre features such as use of personal pronouns and the impact of these on the objective tone.

9.2.3.1 Microplanning (Immediate Planning, Summarising Content, Linking Paragraphs)

As shown by the Table 7.1, Microplanning had the highest number of instances of any category due to the inclusion of Immediate Planning. While the B2 participants reported more instances in the raw data, the C1 participants employed the phase more frequently as a proportion of their overall time (see Table 7.8). Table 7.9 shows that many of the comments made referred either to the paragraph on which the participant was working, consideration of negative or positive points for the content and references to the conclusion. References to the introduction came 24th on the list but this may well be to do with the linear nature of text production in that it was produced early on in the writing process and participants tended to focus on what was yet to be written rather than on what had already been produced.

The C1 participants were also the ones who made most use of the category of Linking Paragraphs and the extracts in 7.13 illustrate how this involved ensuring paragraphs had their own distinct identity (Extract 7.13 a) or else allowed for a better transition to the next paragraph (Extract 7.13 b and c). Extract 9.5 illustrates that a C1 participant saw the explicit linking of ideas in the text as being a development of the process of Text-Level Organisation and that it had a reader-orientated function.

Extract 9.5

<\$2> Yeah I have already linked my thoughts but I <\$E> pause
<\\$E> always try to erm make it easy for the reader to see how my thoughts are linked.

<\$1> Hmm

<\$2> So I just wanted to use correct er linking words.

(Post-interview - Participant 2, C1 <01:06:05>)

Summarising Content, as set out in 7.10 was seen by both B2 and C1 participants as being part of creating the conclusion.

9.2.3.2 Microplanning (Immediate Planning, Summarising Content, Linking Paragraphs) and the models

Microplanning is identified by Field (2004, p. 329) as an intersection between macrolevel decisions about content and organisation and the immediate production of text. Crucially Microplanning involves consideration of the text that has been produced so far with writers needing to consider the status of information and whether it is new or given in the text (Shaw & Weir, 2007, p. 39). Microplanning can be difficult to identify in that ahead of production many of the processes, like those of translation may be automatic and unobservable yet they directly result in the text which is produced (Field, 2004, p. 329). Nevertheless, the data found in the verbal reports suggest that there are different aspects which B2 and C1 writers attended to when producing text.

When Microplanning, the C1 participants made many comments about expressing their ideas clearly or finding the best ways to get these across as depicted in the extracts in 9.6.

Extract 9.6

(a)

<\$2> Er start writing the positive impacts and I'm just trying to think of more <\$=> more of them and er <\$E> pause <\\$E> putting them in order.

(Participant 1, C1 <13:06>)

(b)

Er now because I have written the arguments I am going to use erm I'm trying to <\$E> pause <\\$E> erm <\$E> pause <\\$E> <\$=> to explain er the first argument I used and make clear what I mean by this.

(Participant 11, C1 <09:59>)

(c)

I thought about how to start the conclusion <\$E> pause <\\$E> erm but er without erm telling again er the things that I have talked about <\$=> I have written in the er introduction <\$=> not repeat

(Participant 11, C1 <22:39>)

(d)

<\$2> Okay I was writing down <\$E> pause <\\$E> er my thought about the science and I was trying to find a good way to present it.

(Participant 2, C1 <18:13>)

(e)

\$2> Okay I <\$E> pause <\\$E> wrote and evolved my <\$E> pause <\\$E> erm first erm <\$E> pause <\\$E> thesis so I'm completed it

and I tried to <\$E> pause <\\$E> use er linking word to pass to my next point.

(Participant 2, C1 <29:19>)

(f)

<\$2> I'm developing the first argument erm and I'm actually erm closing the first paragraph+

(Participant 3, C1 <10:12>)

(g)

<\$2> I've er <\$=> I've been trying to er not state my opinion. I've been trying to <\$E> pause <\\$E> <\$=> to <\$E> pause <\\$E> how can I say it erm <\$=> to state what I think er er is <\$=> to state my opinion but not so clearly er as I will do in the conclusion.

(Participant 3, C1 <47:22>)

(h)

I'm thinking of where I'm going with what I'm writing <\$E> pause <\\$E> I mean where it's going to end and what I'm going to write next.

(Participant 6, C1 <12:39>)

Many of the C1 participants' comments in the Immediate Planning category, illustrated by those above as well as those in 7.13 and 9.5 were to do with how to present the information in the essay. This fits with Field's description of microplanning (2004, p. 329). The comments made by B2 participants were more concerned with what the B2 level participants were actually writing rather than what might be the best way to say it, the fit with what had already been written or what was to be produced next. This may link with the observations from the product strand of the study about the lack of range of metadiscourse markers in categories such as Transition Markers as well as the 'simpler' language

used by B2 participants (see 8.6.1.2). As the B2s carry out more knowledge telling (Bereiter & Scardamalia, 1987), they fall back on the language they are most familiar with to carry out interactional and interpersonal functions (Hasselgren, 1994; Carlsen, 2010).

9.2.4.1 Translation (Searching for Lexis)

As has been stated in the previous section as well as in 6.4.1, Translation along with Microplanning can be difficult to examine because much of the phases are internal and may well be automated (Shaw & Weir, 2007, p. 57) and therefore “not susceptible to direct investigation”. As explained in 6.4.2, Searching for Lexis was placed under Translation since it represented the point at which abstract ideas become written text.

The study found that Searching for Lexis showed a significant difference between the B2 and the C1 participants with the C1 writers carrying out more of this category. In addition to this, there also seemed to be a difference in the way the C1s described what they were doing. Extracts 7.14 and Table 7.11 illustrate that there was a concern to come up with words which were ‘appropriate’ or ‘right’. This finding seems to be largely supported by the results from the product strand of the study (see 8.6) where C1 candidates often used a wider range of markers, applied them in different ways and used more complex language.

9.2.4.2 Translation (Searching for Lexis) and the models

Translation is the point at which a learner faces the conflict between what they want to say in the target language and what they are able to say. An example of this can be seen in when participant 2 (level C1) said “so I faced again the problem of er finding er another word for tradition” <43:50>. Participant 3 also made similar comments which are shown in extract 9.7 below:

Extract 9.7

I've actually having some trouble to find <\$=> finding a word so I spent the two minutes doing this <\$E> laughs <\\$E> trying to find the word.

(Participant 3, C1 <23:24>)

Where a participant was unable to come up with a particular word this could result in circumlocution or errors. Participant 2, also reported that she had been searching for a word and came up with one but “I don’t think it’s the er correct one” (<37:59>). The inability to generate a piece of lexis sometimes forced participants to have to rewrite what they were saying, particularly at the B2 level where there were fewer resources for the learners to draw on. Extract 9.8 illustrates this point when a learner stumbled over a fairly simple word.

Extract 9.8

<\$2> I just started to write <\$E> pause <\\$E> my <\$=> the beginning of the sentence, and then I came <\$E> pause <\\$E> erm to the word medicine <\$E> pause <\\$E> and I couldn't remember it so I crossed it out and st <\$=> tried to think of something else to put instead.

<\$1> Um, okay. So you <\$=> you start <\$=> you started writing a sentence couldn't get the word, and then <\$=> and then sort of started again.

<\$2> Yeah.

(Post-interview - Participant 7, B2 <39:21>)

The CEFR suggests that at the C1 level learners will “express him/herself clearly...without having to restrict what he/she wants to say” (Council of Europe, 2001, p. 110). The concerns of the C1 participants over finding the right or appropriate lexis seem to be supported. B2 learners are described in the Framework as having “lexical gaps” (Council of Europe, 2001, p. 112). In the data gathered from the verbal protocols the B2

learners do seem less worried about clarity or whether a particular word or phrase is right as they are with communicating their message. All of this suggests that the CEFR's predictions are accurate. The product strand (see 8.6) also seems to support this in that C1 candidates often applied a wider range of metadiscourse markers, suggesting that they were choosing their language more carefully despite using a lower proportion of metadiscourse markers (see 8.2.2).

9.2.5.1 Monitoring (Monitoring for Unspecified Purposes, Monitoring Content, Monitoring Language)

Neither monitoring as a phase nor any of the categories which make it up showed a statistically significant difference between groups. However, from the comments it can be observed that there were differences in what participants were reporting in Monitoring for Content (Table 7.12) and Language (Table 7.13) as shown in extracts 7.15, 7.16 and 7.17.

Table 9.1 shows the word frequency for the whole Monitoring phase according to what the participants reported. The third word is 'mistakes' and the instances where this was used referred to language errors (three references by C1 participants and two by B2 participants). This mirrors the results in Tables 7.12 and 7.13 where Monitoring Language scores higher and suggests that monitoring for second-language learners is often linguistically rather than content-focussed as predicted by Eysenck and Keane (2010, p. 447). The authors attribute this to the heavy load on working memory during writing and the resulting difficulties in attending to more complex matters of monitoring organisation, discourse, audience or genre.

Table 9.1. *Word frequency for all participants from the Monitoring phase node (N=12).*

Rank	Word	Count	Weighted Percentage (%)
1	essay	6	2.26
2	writing	6	2.26
3	mistakes	5	1.88
4	one	5	1.88
5	read	5	1.88
6	reading	5	1.88
7	write	5	1.88
8	correct	4	1.50
9	order	4	1.50
10	see	4	1.50
11	think	4	1.50
12	words	4	1.50

9.2.5.2 *Monitoring (Monitoring for Unspecified Purposes, Monitoring Content, Monitoring Language) and the models*

Field (2004, p. 330) argues that monitoring is a process which is constantly employed by skilled writers who are considering the macro level of composition (such as issues of purpose and intended audience) and the micro level (effectiveness of the piece of text being constructed and its role in contributing to the macro-level issues). Due to the complexity of the process, Field suggests that only one level of monitoring can be considered at a time.

The findings from the study, while not statistically significant, do suggest patterns to particular levels. C1 participants tended to monitor more for content, which may be linked to the issues already discussed about pre-planning and generating content freeing up space in the working memory. This allowed the C1 participants to consider issues with content and to think about these earlier on. The B2 participants' comments by contrast

(see the extracts in 7.15) were more to do with the evaluation of what they had written and came towards the end of the writing time.

In terms of Monitoring Language, the C1 participants started earlier in the writing time and as explored in Table 7.13 and extracts 7.17 were concerned not only with locating mistakes but also with issues of style (such as repeating a word too often). These results suggest that for the higher level participants, Monitoring was more of an iterative process as predicted by Field (2004, p. 330) and involved some macro-level considerations while the B2 participants behaved in a manner closer to that predicted by Eysenck and Keane (2010) and focussed on surface-level language.

The post-interview question about what participants did and thought about when they had finished writing (Figure 7.3) also appears to suggest that B2 participants were more inclined to assess whether their writing had fulfilled the task, whereas C1 participants had been addressing this throughout composition (Table 7.12).

9.2.6.1 Revising

The Revising phase (and category) returned a statistically significant result (see Table 7.4) which suggests that the C1 participants made more reference to revising than the B2 participants and attended more to this process. The view has long been held that second language learners are not as efficient at revising their texts as L1 writers. Cohen and Cavalcanti (1990) ascribing some of this to the nature of teacher feedback being almost entirely form-focussed and therefore encouraging learners to prioritise such features. Chenoweth and Hayes (2001, p. 94) propose that second language writers are more efficient at making revisions to their texts post-production and Table 7.14 seems to illustrate this for the B2 participants in that the majority of their revisions took place in the final third of their writing time.

Table 9.2 presents word frequency scores for the Revision phase and suggests that both ‘word’ and ‘vocabulary’ are focussed on (a combined total of 4) but that ‘argument’ was also focussed on (ranked third). As this data is drawn largely from C1 participants it suggests that a focus on content in Revision is something which these participants are concerned with.

Table 9.2. *Word frequency scores for the Revision phase node.*

Rank	Word	Count	Weighted Percentage (%)
1	just	4	2.65
2	paragraph	4	2.65
3	something	3	1.99
4	argument	2	1.32
5	evolving	2	1.32
6	fixed	2	1.32
7	make	2	1.32
8	one	2	1.32
9	seconds	2	1.32
10	thought	2	1.32
11	trying	2	1.32
12	use	2	1.32
13	vocabulary	2	1.32
14	words	2	1.32
15	write	2	1.32

9.2.6.2 *Revising and the models*

As Field suggests (2004, p. 330) Revision is connected to Monitoring and particularly to the lexical level as skilled writers make alterations to the text to ensure that the tone is right for the reader. The comments by C1 participants such as those in extracts 7.17 illustrate that at the higher level

participants were indeed reflecting on monitoring and changing the text in order not only to find errors but to improve the text. The study also found that higher level learners were spending longer (in that they attended more in their verbal reports) on revising their texts, a sign of stronger writers.

To sum up, research questions one and two suggest that the C1 participants did more to prepare and plan for their writing and consequently had more resources to devote to issues of appropriate linguistic choices and considering the purpose of their text. B2 participants seem to be more additive in their planning and writing, which may result in the type of cognitive overcrowding discussed in 2.5.2.2. This in turn may result in the use of 'simpler' and familiar linguistic exponents being used by B2 writers.

9.3 Research Question Three:

Is there a difference in the quantity of metadiscourse markers used by candidates of the ESB International ESOL Examinations at levels B2 and C1 of the CEFR?

Research Question Three was intended to investigate whether candidates at the B2 and C1 levels used different quantities of metadiscourse markers in their writing overall. As set out in 4.7 various studies (Burneikaite, 2008; Bax, Nataksuhara & Waller) have concluded that there is often little difference in the overall amount of metadiscourse markers used by candidates at different levels while others have reported differences (Intaraprawat and Steffensen, 1995).

Hypothesis one explored Research Question Three: There is a statistically significant difference between the number of metadiscourse markers used by candidates at B2 and C1. As explored in Table 8.3, there was only a little variation in the means of the length of the B2 and C1 texts and in the amount of metadiscourse used (Table 8.4). In terms of the amount of

metadiscourse as a percentage of the text, there was a reduction in the amount used by the C1 candidates despite their texts being longer. However, this difference in the use of metadiscourse was not statistically significant. This finding concurs with those of Burneikaite (2008) and Bax, Nataksuhara and Waller (Forthcoming) and suggests that metadiscourse is a consistent feature of writer's texts at different levels of proficiency.

The negative finding, together with those from Burneikaite (2008) and Bax, Nataksuhara and Waller (Forthcoming) is useful as it suggests that it is not the presence or absence of metadiscourse markers *per se* which demonstrates a candidate's abilities of discourse competence but the way in which these markers are actually used. This finding contradicts the observation made by Intaraprawat and Steffensen (1995, p. 268) who suggest that the presence of a higher density of metadiscourse markers is an indication of better ability in writing.

Despite the lack of statistical significance, it was observed that the C1 candidates used fewer metadiscourse markers in their writing as a proportion of the total text. This finding was consistent with that of Bax, Nataksuhara and Waller (Forthcoming) who also found an fewer overall metadiscourse markers used by B2 and C1 candidates (a trend which continued at the C2 level). As has already been stated in 4.7, the Bax, Nataksuhara and Waller study was much larger in scale and it is possible that the small-size of the corpus in the current study meant that the difference could not be detected. However, the observed decline in the amount of metadiscourse does lend qualified support to the emerging proposition that higher level learners do reduce the amount of metadiscourse they use.

9.4 Research Question Four:

What are the functions of the metadiscourse markers used by candidates at level B2 and C1?

As stated in 9.3, it was expected that there would be little or no difference in the overall amount of metadiscourse used by the candidates at B1 and B2. However, it was anticipated that there would be differences in the way the different categories were used. Burneikaite (2008, p. 45) identified in her study that L2 writers underused what she termed 'reader-orientated' and 'evaluative' markers, both functions which would fit into Hyland's 'interactional' function (2005, p. 48-54). Bax, Nataksuhara and Waller (Forthcoming) in their study of candidates at B2, C1 and C2 also found that four out of the five 'interpersonal' categories of metadiscourse marker (which overlaps with Hyland's 'interactional' functions, see Table 4.1) returned significant scores. Such results suggest that there would be differences in the ways in which candidates at different levels in the current study would carry out metadiscourse functions. Three hypotheses were proposed and investigated:

- Hypothesis Two: There is a difference in the proportion of interactive metadiscourse markers used between levels B2 and C1;
- Hypothesis Three: There is a difference in the proportion of interactional metadiscourse markers used between levels B2 and C1;
- Hypothesis Four: There are differences in the way individual categories of metadiscourse marker are used between B2 and C1.

The current study was unable to find any statistically significant evidence to support Hypotheses Two or Three or Four regarding the amount of metadiscourse markers used for the interactive or interactional functions. The findings here perhaps should have been anticipated because the data was still being examined from a macro-level in that individual categories were not being examined at this stage. Consequently, the result was similar to that found for Hypothesis One. It is not clear whether features such as the inclusion of high-scoring items across different categories

such as 'and' in the interactional function and 'I' and 'you' in the interactional function could have potentially distorted the overall Figures and prevented any difference from emerging.

Connolly (2007, p. 7) points out, the absence of statistical significance does not necessarily mean that differences cannot be observed but without statistical significance the findings cannot be extended to the wider population with any degree of certainty. In the case of the data for Research Question Four, a difference was observed in means and proportions of metadiscourse markers used for interactive and interactional functions. However, application of the Bonferroni Correction was required in order to avoid type one errors, but it may have resulted in type two errors.

As shown in Table 8.5, the C1 candidates made more use of interactional functions while the B2s used a larger proportion of interactive markers. Shaw and Weir (2007, p. 49) suggest that at B2 candidates begin to move towards the knowledge transforming model of writing (Bereiter & Scardamalia, 1987). While this may be the case it is also likely that the candidates at the B2 level in this study continue to knowledge tell in some aspects of their writing and the slightly higher use of interactive functions may be indicative of this. The C1 participants on the other hand appear to be devoting proportionally more text to the message and managing the interactional features of their writing. Without statistical significance it is not possible to generalise this finding to wider populations with any certainty but the Figures in Table 8.5 do agree with observations in Bax, Nataksuhara and Waller (Forthcoming) and the findings of Burneikaite (2008, p. 45) that non-native speakers underuse the interactional functions.

To conclude, the results for Research Question Four show that we have found that Hypothesis Two, Three cannot demonstrate statistic differences

but that there are patterns of use within the interactive and interactional functions which are consistent with findings from other studies.

Hypothesis Four examined differences in the use of categories of metadiscourse marker. Use of the Mann-Whitney U Test was unable to identify any differences between B2 and C1 candidates in the use of interpersonal or interactional categories when the Bonferroni Correction was applied. However, a more lenient level of significance did suggest differences in the interactional categories of boosters, self-mentions and hedges. The lack of difference between the B2 and C1 candidates in interactive metadiscourse markers may be related to the tendency of non-native writers to overuse 'textual' language such as transition markers (Kennedy, Dudley-Evans, & Thorp, 2001; Hawkey & Barker, 2004, p. 150; Burneikaite, 2008, p. 43). Intaraprawat and Steffensen (1995, p. 269) report Scarcella's finding that native writers use more attention-getting devices and commentaries on the text in their writing than non-native writers. Intaraprawat and Seffensen (1995, p.270) also propose that non-native writers are forced to rely on explicit language to carry out metadiscourse functions such as code-glossing or managing transitions, perhaps more than a native speaker. Viewed from this perspective, it is possible that the reduction in the proportion of metadiscourse markers at C1 is due to increasing sophistication in the way such functions are handled by candidates at the higher level. However, it again must be noted that the differences observed in the use of interactive categories of metadiscourse markers by the B2 and C1 candidates were not statistically significant.

The increased use of boosters by the C1 writers may be indicative of increased confidence in their willingness to support the points that they make in their arguments and may link to the findings from the process stand of the study that C1 participants put more time into generating and organising content than B2 participants. As covered in 4.6.2.2, studies by

Burneikaite (2008) and Morgan (2008) have arrived at differing conclusions as to whether non-native speakers over or under use Boosters. Table 8.15 shows that in this study C1 candidates used a wider range of individual types of booster (22 different types compared to 16 types used by the B2 candidates). However, the C1 participants were using more categories and more markers within the categories which would fit with some of the predictions made by Intaraprawat and Steffensen that “better essays contain a wider range of forms and more of them” (1995, p. 268).

Self-mention also showed some differences between the B2 and C1 participants although not statistically significant when the stricter level of significance was applied. Differences were anticipated by the researcher as one of the most widely-known features of the academic essay is the avoidance of the use of the first-person pronoun, a rule which is an oft-taught one in academic writing. It was expected that this feature would decline in the higher levels as C1 participants would be more aware of the demands of the genre. The CEFR says very little about genre in terms of how awareness of it develops across levels and one of the few references to this is in the illustrative scale for creative writing (Council of Europe, 2001, p. 62) which proposes the following:

- (B2+) can...follow “established conventions of the genre concerned”.
- (C1) “can write...in a...natural style appropriate to the reader in mind”
- (C2) “can write...in a style appropriate to the genre adopted”.

Although only the B2+ and the C2 descriptors mention genre explicitly, presumably the statement from page 36 for C1 (“the discourse skills characterising the previous band continue to be evident”) should be applied. If we presume that the illustrative comments for creative writing

can be applied in part to essay writing it would suggest that C1 candidates should be able to follow established conventions and the avoidance of the first person pronoun and a decline in the number of self-mentions such as the current study has found in the data. The descriptors for general linguistic range in the CEFR (Council of Europe, 2001, p. 110) suggest that by C1 a user “can select an appropriate formulation”, which suggests that learners at this higher level should be able to bear in mind the requirements of the genre and the avoidance of the self-mentions may be indicative of this.

The final category of metadiscourse markers to show a difference, but only in terms of the more lenient level of significance, was hedges. As set out in 4.6.2.2, hedging can convey the degree of confidence on the part of the writer and hedges are often used to avoid making statements which are absolute, acknowledge the existence of alternative viewpoints or to introduce a note of caution where criticism is anticipated (Morgan, 2008, p. 171). Morgan (2008, p.177) and Burneikaitie (2008, p.42) both suggest that learners tend to underuse hedging, a behaviour which Hyland (1994, p. 252) attributed in part to under-representation of the skill in textbooks. Based on the discussion in 4.6.2.2 and the argument that higher level learners should be more sensitive to the demands of genre and the expectations of the reader, it was anticipated that the C1 learners would carry out more hedging in their texts.

The results showed that overall the C1 candidates demonstrated a higher mean for hedging and that they used more examples of the category (see Table 8.11). In terms of individual types of lexical exponent, all but one of those considered (‘in my opinion’) had a higher level of use by the C1 candidates. The findings suggest that in terms of research question four, candidates in the ESB ESOL International Examinations at level C1 carry out more of the interactional functions of boosting and hedging in line with the expectations of the genre (essay writing) while demonstrating more

awareness of some of the established conventions of the genre, such as reducing instances of self-mention.

9.6 Research Question Five:

To what extent do the metadiscourse functions employed by candidates at levels B2 and C1 match the predictions made by the CEFR regarding the development of discourse competence in learners at these levels?

Section 3.7 sets out how the CEFR describes the development of discourse competence in and between the B2 and C1 levels. In order to answer Research Question Five, Hypothesis Five was devised which proposed that there are differences in the way individual metadiscourse types of markers are used by B2 and C1 candidates. There is also some overlap with the Hypotheses One to Four in that the findings from these questions can also be drawn upon to answer Research Question Five.

It has already been shown that although there are no overall statistically significant differences in the numbers or proportions of metadiscourse markers being used by candidates at B2 and C1, nor in the overall numbers of interactive or interactional functions, there are three categories where differences have been observed (see Table 8.8) although not in terms of significance.

The CEFR suggests that there will be differences in terms of the cohesive devices with learners at B2+ using a “variety” (Council of Europe, 2001, p. 35) and presumably this number would extend at the C1 level. At the same time B2 learners will use a “limited” number of cohesive devices. C1 learners will show “controlled use” of both linkers and cohesive devices.

As set out in 4.6.1.1, the category of Transition Markers contains many features which are considered to be linkers and cohesive devices and in

particular connective devices for addition, contrast and causation. Writers such as Burneikaite (2008, p. 43) Kennedy (Kennedy, Dudley-Evans, & Thorp, 2001) and Hawkey and Barker (2004, p. 150) have identified a tendency for language learners to overuse these even at higher levels of proficiency.

The current study did indeed find that the C1 candidates were using a higher number of individual types of Transition Marker (see Table 8.13) although the difference between B2 and C1 was not statistically significant. This suggests that the C1 candidates in this particular study had, as the CEFR scale descriptors suggest, extended the number of cohesive devices and the variety of linking words they use to signal relationships. However, as stated previously, the overall difference in the number of different types of transition markers used was not statistically significant so this cannot be applied to other populations taking timed tests of essay writing at the B2 and C1 levels though it may be a feature which further research with a larger study could investigate. These differences are illustrated in Table 8.14. This Table highlights a number of features which are consistent with CEFR scale descriptors. First of all, C1 candidates made less use of the causative linker 'because' than B2 candidates. 'Because' is very high-frequency and an early piece of language that learners are likely to be exposed to and make use of. For example, The English Profile (Cambridge English Language Assessment et al., 2014) project identifies 'because' as being an A1 piece of language. Carlsen (2010, p. 201) also identified the Norwegian equivalent of 'because' as a piece of very high frequency language and identified it as one of her 'lexical teddy bears'.

Both 'because' and 'because of' have more instances of use in the B2 data (note that they were counted together in the statistical analysis). In terms of markers for causation at the C1 level it is the candidates' use of different contrastive linkers such as 'although', 'nevertheless' and

‘whereas’ which stands out in particular, with the C1s demonstrating a wider range of lexical exponents for this function. The mean for ‘but’ also drops at the C1 level, another high-frequency piece of language according to both Carlsen (2010) and English Profile (Cambridge English Language Assessment et al, 2014) but not enough for the difference to be statistically significant.

In terms of connectors with an additive function, ‘besides’ and ‘in addition’ were both used more by C1 candidates along with increased use of ‘furthermore’, ‘moreover’ and a decline in the use of ‘also’. It therefore appears that the CEFR’s assertion that the range of linking devices and cohesive devices (in terms of connectors) does extend at the C1 level and that there is increasing complexity in the linking language and cohesive devices which learners at the higher level choose to use in their writing.

Linked to this is a comment from one of the C1 participants in the process strand who reported that in the pre-planning stages they considered the linking language that they would use.

Extract 9.8

And then the linking words that is appropriate every time.
(Post-interview – Participant 11, C1 <35:30>)

Other C1 participants also showed concern about choosing language which demonstrated their level and was appropriate (e.g. Participant 1 <30:00>, Participant 3 <56:06>) and as shown in Extracts 7.13 and 7.14. It is possible that by thinking about linking language at the planning stage or by taking the time to search for more advanced lexis (which may include connectors) writers at C1 differentiate their writing from that of learners at lower levels of proficiency. It is also possible that the knowledge telling (Bereiter & Scardamalia, 1987) approach to writing of B2 candidates, with ideas being generated as they write leads to a crowding of the working

memory forcing these writers to rely on familiar linking devices and structures.

In terms of organising their texts and showing control of organisational patterns (Council of Europe, 2001, p. 36) C1 candidates used fewer items in the sequencing or label stages categories (although the difference was not statistically significant). It is interesting that the B2 participants used more of the sequencing marker 'secondly' as it suggests that although C1 writers were using 'firstly' they felt sufficiently in control of their texts not to need to structure their texts with an ordinal list. This again may suggest more control and greater range in terms of linguistic exponents for metadiscourse and a willingness to move away from high-frequency markers.

Self-mention was another category where both 'we' and 'us' were used more by B2 learners in terms of mean. As discussed in 9.5 above, the lower overall figure by C1 participants for use of Self-mention may be indicative of deepening genre awareness with the writers seeking a more objective tone.

In conclusion, to answer Research Question five, it appears that the CEFR is correct in its predictions that the variety of linkers and connectives will extend. It also seems that there may be more awareness of the expectation of the readers in the choice of linguistic exponents made by higher-level writers as well as in their recognition of some features of the genre such as objective tone.

9.7 Conclusion

Bringing together the results from both strands of the study suggests that differences can be observed in the writing process and products of candidates at B2 and C1. B2 candidates appear to be more concerned with producing text than planning which may result in the use of more

familiar language to carry out metadiscourse functions. C1 candidates appear to focus on the task requirements and purpose and use a range of language to carry out metadiscourse functions. Many of the CEFR's descriptions of B2 and C1 learners appear to be supported. The small-scale nature of the study means that there is a risk of some type two error in the reporting of statistical differences but observation of the data suggests patterns which might be detectable in a larger study. Limitations will be discussed in more detail in 10.3.

Chapter Ten: Conclusions

10.1 Introduction

This chapter begins by discussing the conclusions drawn from both strands of the study before considering the limitations on the results reviewed in the previous chapter in terms of the methods used and the findings. The chapter will then discuss the implications of the study for language testing and the ESB ESOL International English Examinations in particular before looking at potential implications for language teaching. The final section will consider areas for further research.

10.2 Conclusion: The aims of the study

The project reported in this study had the aim of answering two questions, these were:

1. To what extent is **cognitive validity demonstrated** in the cognitive phases that candidates carry out while producing scripts at levels B2 and C1 in the English Speaking Board ESOL International Examinations?
2. What is the role of discourse competence in deciding whether a script is classified as being level B2 and C1 of the Common European Framework of Reference for Languages (CEFR) in candidate scripts from the ESB ESOL International Examinations?

In terms of the first aim, the results for research question one have shown that the essay writing tasks did elicit a wide range of processes at both B2 and C1 and that all of Field's (2004) cognitive phases were represented in the verbal reports of participants. Statistical differences were found in the way B2 and C1 participants selected the lexis they used and revised their texts. Other differences in behaviour were also identified within the cohort in terms of how and when participants at the different levels planned and

generated content, but these would have to be explored further before any claims could be made regarding the wider population of test-takers. The results from the product strand of the project also indicated a wider range of metadiscourse markers were used for functions such as Transition markers and that there was a reduction in some 'simpler' markers such as 'because' and 'but'. This may indicate that higher-level writers made more effort to use appropriate and more varied markers within their discourse and were able to do so due to having carried out their content generation and planning earlier on in the writing process.

Research question two, which also related to the first aim of the study, also found that there was evidence for many of the features identified by different models of cognitive processing. The C1 participants did seem to be more aware of the audience and select lexis accordingly and think about the impact of their text when revising. The verbal reports also suggested that the CEFR's predictions about the continuing development of discourse competence at the C1 level were correct. However, the writing tasks used in the ESB ESOL International tests would almost certainly benefit from being more clearly specified in terms of intended audience and text purpose to allow the C1 participants to better demonstrate these skills.

In terms of the second research aim, research question three identified no overall difference in the amount of metadiscourse markers used overall, which was predicted by Burneikaite (2008). Nor did research question four find differences between interactive and interactional categories. However, the process strand of the study suggested that C1 candidates were more concerned about the communication of the message and the expectations of the reader, as the CEFR predicts and the qualitative exploration of the product data appears to support this due to the greater variety of forms used. The review of individual exponent types of metadiscourse markers in research question five also suggests that the CEFR is correct to predict

a more varied range in some areas, particularly in the use of transition markers. The results suggest that discourse competence and the use of such markers and the reduction of use in high-frequency exponents could be considered for use in writing criteria at the C1 level. As suggested above, better specification of the audience and purpose of the text would almost certainly assist in giving higher-level candidates the opportunity to better demonstrate these skills.

10.3 Limitations of the study

While every attempt has been made to produce a carefully thought-out and principled study, no piece of academic work is without limitations. While limitations represent defects in the study, they also present potential opportunities for further research.

The first issue which affects both the product and the process strands is the number of samples in both. As set out in 6.3.1 a total of twelve participants were used for the process strand of the study with six participants at B2 and six at C1. While this is a larger number than in some other studies such as Plakans (2009, p. 567), it remains a comparatively small number. To some extent this limitation is a constraint on most verbal report studies due to the time consuming nature of the transcription stage (Green, 1998, p. 50) but nevertheless it does mean that only a relatively small number of learners were able to be sampled for the process strand.

The number of samples in the product strand of the study was also a problem. Other studies have often used quite small sets of texts. Intaraprawat & Steffensen (1995, p. 95) examined a total of twelve essays, six 'good' and six 'poor' while Burneikaite (2008, p.40) looked at 40 MA dissertations (20 native speaker and 20 non-native speaker). Those studies which have used larger numbers such as Carlsen (2010, p. 198) or Bax, Nataksuhara and Waller (Forthcoming) have been unable to control

for task or nationality thereby introducing two potentially quite far-reaching intervening variables into their results. As described in 6.5.8 the issue was with the comparatively small number of candidates in Greece who take the C1 examination. Due to the financial and time constraints on learners, the *frontistiria* English schools often encourage those students who have successfully completed the B2 level to jump to the C2 level. The issue of learners skipping C1 the level is compounded by the fact that very few of the learners who sit the C1 examination actually choose the essay question and prefer the letter option. This letter option was not chosen for the study due to the author's doubts about its suitability as a C1 level format. The final issue was that of those who do take the C1 exam, and opt for the essay, only learners who had passed the whole examination (and in all sections) were selected for the corpus. While this did result in a pool of 90 scripts, this was reduced by the independent rating process described in 5.4.2.6. In the end thirty scripts at each level (B2 and C1) were chosen as this was felt to be the minimum size at which statistical data could be analysed (Hatch & Lazaraton, 1990). This smaller sized corpus and the lack of normal distribution of the data necessitated non-parametric tests and the use of a more stringent test of statistical significance using the Bonferroni Correction in order to avoid type one errors (false claims of significance) it may also result in type two errors (failing to identify significance) (Pallant, 2005, p. 200).

Connected to the issue of the number of samples in both the process and the product strands of the study is the issue of multi-dimensionality of CEFR level (de Jong, 2009) and the issue of how well each participant (in the process study) and candidate script (in the product study) fit with the level which they were representing. While all of the participants in the verbal reports had achieved the level they were representing in all skills in the Cambridge ESOL examinations, some features of their writing appeared closer to the adjacent level. For example, participant 10 who was representing B2 appeared to engage more in pre-planning and the

types of behaviours which appeared more in the C1 participants, while participant 6 (C1) seemed to exhibit some behaviour that was more typical of B2-level participants.

With the candidate scripts in the product strand, independent raters sometimes allocated higher or lower grades to the different bands of candidates, so that even if they identified someone as a B2, the rater felt that the candidates were demonstrating some features at B1+ or C1. What this emphasises is that learner development through the CEFR is not linear or lock-step. Different features develop at different times in individuals. It also suggests that rather than simply looking at features of B2 learners or features of C1 learners, we ought to consider the range of performances *within* the level. Perhaps some of the difficulty in identifying statistical significance is due to the variance of performance within the levels which may be to do with the way the C1 corpus was selected. However, since there is a similar variety in the performances of the B2 participants, this range may simply reflect the multi-dimensionality of the writing skill within the levels and the lack of a clear cut-off between the B2/B2+ and C1 levels.

A further limitation on the study is that only participants and samples from one nationality was used and only one genre was considered (essay writing) and only one type of task (advantage/disadvantage essay). As described in Chapter Six, these decisions were taken partly for practical reasons such as obtaining sufficient sample sizes but also in an attempt to control the variables in the study. As set out above, larger studies such as those by Carlsen (2010, p. 198) or Bax, Nataksuhara and Waller (Forthcoming) have often had to take learners from a range of nationalities as well as a range of different task types. In the current study it was hoped to control these two variables so that although only one task type was used a number of different variations on it could be accepted. This does open the question of whether candidates from different nationalities would

behave differently both in terms of composition (in the process study) and product (in the product strand of the study). In terms of genre it is also possible that a different text type (such as a letter or report task) would result in eliciting different writing processes or metadiscourse markers. It is also possible that an essay which required learners to use a different rhetorical pattern such as compare/contrast, agree/disagree or problem/solution might well produce different results.

Related to this issue of the task type is the fact that there were no differences other than text length for the B2 and C1 participants in the process strand of the study. As set out in 6.5.4, the task was chosen since the advantage/disadvantage argument was felt appropriate for both levels and the addition of the 'why' element of the question was deemed by task raters to provide the additional challenge for C1 participants. The task was typical of those used in the ESB tests, however, as has been argued, the impact of the set task is a very important consideration and it is possible that the fact that C1 candidates were required to write more may account for some of the differences.

Another important limitation on the study is the age range of the candidates who take the B2 level examination in particular. Since some candidates may be as young as fourteen years old, arguably the cognitive processes which they can bring to bear on the task may be very different. They may not have developed competent writing skills in their L1 by this stage so it would be highly unlikely that they would be able to apply such skills in their L2 writing.

Finally, the lack of specification of audience is a clear deficiency in the set writing task. In order to be consistent with the ESB ESOL International Examinations and the aim of the project to investigate these tests, this lack of explicit specification was extended to both the process and product writing tasks but the feeling of the author is that it is unhelpful and unfair to

candidates not to specify the reader because it means that the writer is placed in the position of having to guess. This in turn impacts on the decisions made by the writer about how to acknowledge or involve the reader. As stated, while the author acknowledges that the audience ought to have been specified, the decision was taken to mirror the actual tasks used in the examinations.

10.4 Implications for language testing, the ESB ESOL International English Language Examinations, teaching and further research

As discussed above, the results from this study have limitations on them but they do suggest some implications for language testing and for the development of the ESB ESOL international English language examinations in particular. Firstly, metadiscourse functions such as hedging and the use of self-mention could be used in writing assessment criteria as indicators of level. C1 candidates appear to be more aware of “established conventions” (Council of Europe, 2001, p. 62) which means that these aspects of genre could be more clearly identified in a criteria. The range and variety of transition markers in terms of connectors could also be identified as well as some guidance around how high-frequency metadiscourse markers could be expected to be less-used. Of course any such criteria would need developing and piloting but the findings in this study suggest that there are features which could be considered.

Further research in this area could look at different written task types such as reports, e-mails, letters and narratives. Different rhetorical patterns should also be considered so that common patterns such as problem-solution are investigated to see whether they generate different metadiscourse.

Carlsen’s observation (2010, p. 203) that the CEFR may require revisiting in the light of the range of connectors used by learners at lower levels appears to be supported by this study. B2 candidates in the product

strand of the study use more than a 'limited' range of linking devices and if evidence were generated from studies in other languages (in addition to Norwegian and English) then this might necessitate some rewording of these descriptors. However, an area that the author feels merits further investigation is the multi-dimensional aspect of CEFR levels and in particular the range of performances intra-level. This study, both the process and the product parts, could be replicated using B1+ candidates, as well as candidates who have achieved low-passes at the B2 and C1 levels and those who have achieved high passes. This study could then look at how performance and products develop within and between levels. Such specification would be invaluable to testers, teachers and course and materials developers as levels such as B2 are potentially enormous in scope and even in this limited study overlaps between higher and lower level performance have been found within the samples.

A further implication, as discussed in 10.2, is that the ESB ESOL Examinations ought to provide more explicit information about the intended audience for the essay so as to give more validity to any criteria linked to text and task purpose since without this specification candidates are left to draw on their own resources. Clearer specification might offer more scope to stronger writers at B2 and C1-level writers to tailor their interactional metadiscourse more towards the intended reader.

Finally, in terms of language teaching, the results from the verbal reports suggest that teaching learners how to plan their essays, particularly at the B2 level, needs to go beyond question analysis. Johnson (2012) asserts that planning may have little impact on lower level learners and from this study it appears that sometimes this planning simply involves writing down the key words from the question. C1 candidates by way of contrast seem to invest more time in generating content, coming up with key lexis, considering linkers they might use and organising their ideas before they start writing and in so doing perhaps free themselves to focus more on the

language that they use in the task. Perhaps it might be possible to use a variation on the verbal reporting method used in this study with learners in the classroom to provide them with insights into how they write and the impact of attempting to generate content and language simultaneously. All of the participants in the process study said that they had found the activity helpful and that it would make them think about how they approached writing tasks in the future.

References

- Adolphs, S. (2008). *Corpus and Context: Investigating Pragmatic Functions in Spoken Discourse*. London: John Benjamins Publishing Company.
- Angouri, J. (2010). Combining methods in linguistic research. In L. Litosseliti, *Research Methods in Linguistics* (p. 29-45). London: Continuum.

- Ashwell, T. (2000). Patterns of teacher response to student writing in a multiple-draft composition classroom: Is content feedback followed by form feedback the best method? *Journal of Second Language Writing*, 9(3), 227-257.
- Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
- Bachman, L., & Palmer, A. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.
- Baddeley, A., & Hitch, G. (1974). Working memory. In G. Bower, *The Psychology of Learning and Motivation: Advances in Research and Theory*, vol. 8. New York: Academic Press.
- Badger, R., & White, G. (2000). A process genre approach to teaching writing. *ELT Journal*, 54(2), 153-160. doi:10.1093/elt/54.2.153.
- Baker, P. (2010). Corpus methods in linguistics. In L. Litosseliti, *Research Methods in Linguistics* (p. 93-113). London: Continuum International Publishing Group.
- Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing* 28 (1), 51-75.
- Bax, S., Nataksuhara, F., & Waller, D. (Forthcoming). *Researching metadiscourse markers in candidates' writing at Cambridge FCE, CAE and CPE levels*. Cambridge: Cambridge University Press.
- Bereiter, C., & Scardamalia, M. (1987). *The Psychology of Written Composition*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Bhatia, V. (1993). *Analysing Genre: Language Use in Professional Settings*. London: Longman.

- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243-257.
- Bond, T., & Fox, C. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences (2nd Edition)*. London: Lawrence Erlbaum Associates.
- Brooks, A., & Grundy, P. (1991). *Beginning to Write: Writing activities for Elementary and Intermediate Learners*. Cambridge: Cambridge University Press.
- Brown, A. (2004). Legibility and the ratings of second-language writing. In M. Milanovic, & C. Weir, *European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference July 2001: Studies in Language Testing* 18 (p. 117-125). Cambridge: Cambridge University Press.
- Brown, D. (2002). English language teaching in the “post-method” era: toward better diagnosis, treatment and assessment. In J. Richards, & W. Renandya, *Methodology in Language Teaching: An Anthology of Current Practice* (p. 9-18). Cambridge: Cambridge University Press.
- Brown, H. (2004). *Language Assessment: Principles and Classroom Practice*. New York: Pearson.
- Brown, J. D., & Rodgers, T. S. (2002). *Doing Second Language Research*. Oxford: Oxford University Press.
- Burneikaite, N. (2008). Metadiscourse in linguistics masters' theses in English L1 and L2. *Kalbotyra*, 59(3), 38-47.
- Cambridge English Language Assessment; Cambridge University Press; The British Council; The University of Cambridge; The University of Bedfordshire; English UK. (2014). *English Profile: CEFR for*

English. Retrieved August 14 , 2014, from
<http://www.englishprofile.org/>.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. Richards, & R. Schmidt, *Language and Communication* (p. 2-27). New York: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative competences to second language teaching and testing. *Applied Linguistics*, 1(1), 1-47. Retrieved 06 2010, 17, from
<http://applied.oxfordjournals.org>.

Candlin, C., & Hyland, K. (1999). *Writing: Texts, Processes and Practices*. London: Longman.

Carlsen, C. (2010). Discourse connectives across CEFR-levels: a corpus based study. In I. Bartning, M. Martin, & I. Vedder, *Communication Proficiency and Language Development: Intersections between SLA and Language Testing Research* (p. 191-210). European Second Language Association.

Carmen Campoy-Cubillio, M., Belles-Fortuno, B., & Lluïsa Gea-Valor, M. (2010). *Corpus-Based Approaches to English Language Teaching*. London: Continuum International Publishing Group.

Celce-Murcia, M., Dornyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics*, 6(2), 5-35.

Chan, S.H.C. (2011). Research note: demonstrating cognitive validity and face validity of PTE Academic Writing items Summarise Written Text and Write Essay. Retrieved 15/06/2015 from
pearsonpte.com/research/research-summaries-notes/

- Chenoweth, N. A., & Hayes, J. R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication*, 18(1), 80-98.
doi:10.1177/0741088301018001004
- Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, Mass.: MIT Press.
- Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *Canadian Modern Language Review*, 59(3), 393-423.
- Cobb, T. (2014). *Compleat Lexical Tutor*. www.lextutor.ca.
- Cohen, A., & Cavalcanti, M. (1990). Feedback on compositions: Teacher and student verbal reports. In B. Kroll, *Second Language Writing: Research Insights for the Classroom* (p. 155-177). Cambridge: Cambridge University Press.
- Cohen, L., Manion, L., & Morrison, K. (2000). *Research Methods in Education (5th Edition)*. London: Routledge.
- Connolly, P. (2007). *Quantitative Data Analysis in Education: A Critical Introduction Using SPSS*. Abdingdon: Routledge.
- Conway, M., & Holmes, E. (2005). Autobiographical memory and the working self. In N. Braisby, & A. Gellatly, *Cognitive Psychology* (p. 507-543). Oxford: Oxford University Press.
- Council of Europe. (2001). *The Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press.
- Council of Europe. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching and Assessment (CEFR) - A Manual*. Strasbourg: Language Policy Division.

- Crabtree, B., & Miller, W. (1999). *Doing Qualitative Research (2nd Edition)*. London: Sage Publications.
- Crismore, A., Martkkanen, R., & Steffensen, M. (1993). Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Written Communication*, 10(1), 39-71.
- Cronbach, L. J. (1971). Test Validation. In R. Thorndike, *Educational Measurement (2nd edition)*. (p. 443-507). Washington D.C.: American Council on Education.
- Davies, M. (2004). *Brigham Young University / British National Corpus Interface*. Retrieved from BYU-BNC: <http://corpus.byu.edu/bnc/>.
- de Jong, J. (2009, June 8). *Unwarranted claims about CEF alignment by some international English tests*. Retrieved May 12, 2014, from www.ealta.eu.org:
http://www.ealta.eu.org/conference/2009/docs/friday/John_deJong.pdf.
- Dornyei, Z. (2007). *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.
- Ellis, R., & Yuan, P. (2004). The effects of planning on fluency, complexity, and accuracy in second language writing. *Studies in Second Language Acquisition*, 26, 59-84.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review* 87(3), 215-251.
- Ericsson, K. A. (2002). Towards a procedure for eliciting verbal expression of non-verbal experience without reactivity: interpreting the verbal overshadowing effect within the theoretical framework for protocol analysis. *Applied Cognitive Psychology* 16, 981-987.
- Ericsson, K., & Simon, H. (1984). *Protocol Analysis: Verbal Reports as Data*. London: The Massachusetts Institution of Technology Press.

- Eysenck, M. (1993). *Principles of Cognitive Psychology*. Hove: Lawrence Erlbaum Associates Ltd.
- Eysenck, M., & Keane, M. (2010). *Cognitive Psychology (Sixth edition)*. Hove: Psychology Press.
- Field, J. (2004). *Psycholinguistics: The Key Concepts*. Retrieved 03 07, 2010, from Routledge Student Net Library:
www.netlibrary.com/reader
- Firth, J. R. (1950). Personality and language in society. In *The Sociological Review*, 42(1), 37-52.
- Fraser, B. (1996). Pragmatic markers. *Pragmatics*, 6(2), 167-190.
- Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.
- Gathercole, S. E., & Baddeley, A. D. (1993). *Working Memory and Language*. Hove: Lawrence Erlbaum Associates Publishers.
- Gibson, S. (2008). Reading aloud: A useful learning tool? *ELT Journal*, 62(1), 29-36. doi:10.1093/elt/ccm075.
- Grabe, W., & Kaplan, R. (1996). *Theory and Practice of Writing*. Harlow: Pearson Education Limited.
- Green, A. (1998). *Verbal Protocol Analysis in Language Testing Research: A Handbook*. Cambridge: Cambridge University Press.
- Grotjahn, R. (1987). On the methodological basis of introspective methods. In C. Faerch, & G. Kasper, *Introspection in Second Language Research*. Clevedon: Multilingual Matters.
- Halliday, M. (1973). *Explorations in the Function of Language*. New York: Elsevier.

- Halliday, M. (1985). *An Introduction to Functional Grammar*. London: Edward Arnold.
- Halliday, M. (1994). *An Introduction to Functional Grammar (2nd edition)*. London: Edward Arnold.
- Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the way Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, 4, 237-258.
- Hatch, E.M., & Lazaraton, A. (1991). *The Research Manual: Design and Statistics for Applied Linguistics*. New York: Newbury House.
- Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, 9, 122-159.
- Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy, & S. Ransdell (Eds.), *The Science of Writing*. (p. 3-30). NJ: Lawrence Erlbaum Associates.
- Hayes, J., & Flower, L. (1980). Identifying the organization of writing processes. In L. Gregg, & E. Steinberg, *Cognitive processes in writing* (p. 31-50). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hinds, J. (1987). Reader versus writer responsibility: A new typology. In U. Connor, & R. Kaplan, *Writing across Languages: Analysis of L2 Text* (p. 141-152). Reading, MA: Addison-Wesley.
- Hitch, G. (2005). Working memory. In N. Braisby, & A. Gellantly, *Cognitive Psychology* (p. 307-343). Oxford: Oxford University Press.
- Hyland, K. (1994). Hedging in academic writing and EAP textbooks. *English for Specific Purposes*, 13(3), 230-256.
- Hyland, K. (1999). Disciplinary discourses: writer stance in research articles. In C. Candlin, & K. Hyland, *Writing: Texts, Processes and Practices*. London: Longman.

- Hyland, K. (2002). *Teaching and Researching Writing*. Harlow: Pearson Education.
- Hyland, K. (2004). *Disciplinary Discourses*. Michigan: University of Michigan Press.
- Hyland, K. (2005). *Metadiscourse: Exploring Interaction in Writing*. London: Continuum.
- Hyland, K., & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied Linguistics*, 25(2), 156-177.
- Hymes, D. (1972). On communicative competence. In J. Pride, & J. Holmes, (Eds.), *Sociolinguistics* (p. 269-293). Harmondsworth: Penguin.
- IBM Corp. (2013). *IBM SPSS Statistics Version 22*.
- IELTS. (2012). *IELTS/Researchers/Common European Framework*. Retrieved 10 21, 2013, from http://www.ielts.org/researchers/common_european_framework.aspx.
- Intaraprawat, P., & Steffensen, M. (1995). The use of metadiscourse in good and poor ESL essays. *Journal of Second Language Writing*, 4(3), 253-272.
- Johnson, K., Kim, M., Ya-Fang, L., Nava, A., Perkins, D., Smith, A., Lu, W. (2008). A step forward: investigating expertise in materials evaluation. *ELT Journal*, 62(2), 157-163.
- Johnson, M.D., Mercado, L., & Acevedo, A. (2012). The effect of planning sub-processes on L2 writing fluency, grammatical complexity, and lexical complexity. *Journal of Second Language Writing*, 21, 264-282. doi: 10.1016/j.jslw.2012.05.011

- Johnson, M.D. (2014). Does planning really help?: Effectiveness of planning in L2 writing. *Journal of Second Language Teaching and Research*. 3(1), 107-118.
- Jones, E. (2014, 08 7). *Doctor Who series 8, episode 1 review: Peter Capaldi's Doctor may just be the best yet*. Retrieved 08 09, 2014, from The Independent Newspaper:
<http://www.independent.co.uk/arts-entertainment/tv/reviews/doctor-who-deep-breath-bfi-southbank-tv-review-peter-capaldi-doctor-may-just-be-the-best-yet-9655675.html>
- Kane, M. (2006). Validation. In R.L. Brennan (Ed.), *Educational Measurement* (p. 17-64). Washington DC: American Council on Education/Praeger.
- Kellogg, R. T. (1999). *The Psychology of Writing*. New York: Oxford University Press.
- Kennedy, C., Dudley-Evans, T., & Thorp, D. (2001). *Investigation of Linguistic Output of Writing Task Two*. British Council.
- Khajavy, G., Asadpour, S., & Yousefi, A. (2012). A comparative analysis of interactive metadiscourse features in discussion sections of research articles written in English and Persian. *International Journal of Linguistics*, 4(2), 147-149. Retrieved 4 24, 2013, from <http://www.macrothink.org/journal/index.php/ijl/article/view/1767/pdf>
- Khezrlou, S. (2012). The relationship between cognitive and metacognitive strategies, age, and level of education. *The Reading Matrix*, 12(1), 50 - 61. Retrieved 02 07 2015 from http://www.readingmatrix.com/articles/april_2012/khezrlou.pdf
- Kvale, S. (1996). *Interviews*. London: Sage Publications.
- Lakoff, G. (1972). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Chicago Linguistic Society Papers*, 8, 182-228.

- Lashley, K. S. (1923). The behavioristic interpretation of consciousness II. *Psychological Review*, 30, 329-353.
- Lautamatti, L. (1978). Observations on the development of the topic in simplified discourse. In V. Kohonen, & N. Enkvist, *Text Linguistics, Cognitive Learning and Language Teaching* (p. 71-104). Turku: University of Turku Publications.
- LeCompte, M., & Preissle, J. (1993). *Ethnography and Qualitative Design in Educational Research (2nd Edition)*. London: Academic Press Limited.
- Levon, E. (2010). Organising and processing your data: The nuts and bolts of quantitative analysis. In L. Litosselti, *Research Methods in Linguistics* (p. 68-92). London: Continuum.
- Lim, G. (2012). Developing and validating a mark scheme for writing. *Cambridge ESOL: Research Notes*, 49, 6-10.
- Linacre, J.M.(1989). *Many Facet Rasch Measurement*. Mesa Press.
- Linacre, J. M. (2008). *Facets Rasch model computer program* [Software manual]. Chicago: Winsteps.com.
- Luzon, M., Campoy, M., Sanchez, M., & Salazar, P. (2007). Spoken corpora: New perspectives in oral language use and teaching. In M. Campoy, & M. Luzon, *Spoken Corpora in Applied Linguistics* (p. 3-26). Bern: Peter Lang.
- MacArthur, C., Graham, S., & Fitzgerald, J. (2006). *Handbook of Writing Research*. New York: The Guilford Press.
- Manchon, R.M. & Roca de Larios, J. (2007). on the temporal nature of planning in L1 and L2 composing. *Language Learning*, 57, 549 - 593.
- Martin, J., & Rose, D. (2003). *Working with Discourse: Meaning Beyond the Clause*. London: Continuum.

- McCarthy, M., & Carter, R. (1994). *Language as Discourse: Perspectives for Language Teaching*. London: Longman.
- McEnery, T., & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- McIntosh, A., Halliday, M. A. K & Strivens, P., (1964). *The Linguistic Sciences and Language Teaching*. London: Longman.
- McNamara, T. (1995). Modelling performance: Opening Pandora's box. *Applied Linguistics*, 16(2), 159-179. doi:10.1093/applin/16.2.159.
- Messick, S. (1989). Validity. In R. Linn, *Educational Measurement (3rd edition.)* (p. 13-104). New York: Macmillan.
- Michkan, P., Slater, S. & Gibson, C. (2000). Study of response validity of the IELTS writing subtest. *IELTS Research Reports 3* (2000). Retrieved 29/06/2015 from http://www.ielts.org/researchers/research/ielts_writing_test.aspx
- Microsoft Corporation. (2013). *Microsoft Excel 2013*. Redmond WA.
- Microsoft Corporation. (2013). *Microsoft Word 2013*. Redmond WA.
- Morgan, B. (2008). The space between Yes and No: how Italian students qualify and boost their statements. In M. Pawlak (Ed.), *Investigating English Language Learning and Teaching* (p. 267-278). Poznań: Adam Mickiewicz University.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231-59.
- Nystrand, M. (1982). *What Writers Know: The Language, Process, and Structure of Written Discourse*. New York: Academic Press.

- O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.
- Ong, J., & Zhang, L. (2010). Effects of task complexity on the fluency and the lexical complexity in EFL students argumentative writing. *Journal of Second Language Writing*, 19, 218-233.
- Pallant, J. (2005). *SPSS Survival Manual*, 2nd Edition. Maidenhead: Open University Press and McGraw-Hill Education.
- Phakiti, A. (2006). Modeling cognitive and metacognitive strategies and their relationships to EFL reading test performance. *Melbourne Papers on Language Testing*, 2006, 53 - 96. Retrieved 01 07 2015 from http://www.http://sydney.edu.au/education_social_work/research/publications/resources/phakiti.pdf
- Plakans, L. (2008). Comparing composition processes in writing-only and reading-to-write test tasks. *Assessing Writing*, 13(2), 111-129.
- Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, 26(4), 561-587. doi:10.1177/026553220940192
- Purpura, J.E. (1999). *Learner Strategy Use and Performance on Language Tests: A Structural Equation Modeling Approach*. Cambridge: Cambridge University Press.
- Qi, X., & Liu, L. (2007). Differences between reader/writer responsible languages reflected in EFL learners' writing. *Intercultural Communication Studies*, 16(3), 148-159. Retrieved 5 15, 2013, from <http://www.uri.edu/iaics/content/2007v16n3/12%20Xiukun%20Qi%20&%20Lida%20Liu.pdf>

- QSR International Pty Ltd. (2014). *NVivo 10 for Windows*.
- Rasinger, S. (2010). Quantitative methods: concepts, frameworks and issues. In L. Litosseliti, *Research Methods in Linguistics* (p. 49-67). London: Continuum International Publishing.
- Richards, J., & Rodgers, T. (2001). *Approaches and Methods in Language Teaching*. Cambridge: Cambridge University Press.
- Richards, J., Platt, J., & Weber, H. (1985). *Longman Dictionary of Applied Linguistics*. London: Longman.
- Richards, K. (2003). *Qualitative Inquiry in TESOL*. Hampshire: Palgrave Macmillan.
- Roca de Larios, J., Manchon, R.M., Murphy, L. & Marin, J. (2008). The foreign language writer's strategic behaviour in the allocation of time to writing processes. *The Modern Language Journal*, 90, 100-114.
- Robson, C. (2002). *Real World Research*. London: Blackwell Publishing.
- Ruch, G. (1924). *The Improvement of the Written Examination*. Chicago: Scott, Foresman and Company.
- Schiffrin, D. (1994). *Approaches to Discourse*. London: Blackwell.
- Schiffren, D., Tannen, D., & Hamilton, E. (2001). *The Handbook of Discourse Analysis*. Oxford: Blackwell Publishing.
- Shaw, S., & Weir, C. (2007). *Studies in Language Testing 26: Examining Writing - Research and Practice in Assessing Second Language Writing*. Cambridge: Cambridge University Press.
- Sinclair, J. (1991). *Corpus, Concordance and Collocation*. Oxford: Oxford University Press.

- Sireci, S. (2007). On validity theory and test validation. *Educational Researcher*, 36(8), 477-481. doi:10.3102/0013189X07311609.
- Skulstad, A. (2004). The use of metadiscourse in introductory sections of a new genre. *International Journal of Applied Linguistics*, 15(1), 71-86.
- Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.
- Swales, J. M. (2004). *Research Genres*. Cambridge: Cambridge University Press.
- Thomas, S., & Hawes, T. (1994). Reporting verbs in medical journal articles. *English for Specific Purposes*, 13, 129-148.
- Thompson, G. (1996). Some misconceptions about communicative language teaching. *ELT Journal*, 50(1), 9-15. doi:10.1093/elt/50.1.9
- Van Ek, J., & Alexander, L. (1979). *The Threshold Level in a European Unit/Credit System for Modern Language Learning by Adults*. Strasbourg: Council of Europe.
- Vande Kopple, W. (1985). Some exploratory discourse on metadiscourse. *College Composition and Communication*, 82-93.
- Vergaro, C. (2005). 'Dear Sirs, I hope you will find this information useful': discourse strategies in Italian and English 'For Your Information' (FYI) letters. *Discourse Studies*, 7(1), 109-135. doi:10.1177/1461445605048769
- Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.
- Weir, C. (1988). *Communicative Language Testing*. Exeter: University of Exeter Press.

- Weir, C. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Hampshire: Palgrave Macmillan.
- Weir, C. (2005a). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22(3), 1-20.
- Weir, C., O'Sullivan, B., Yan, J., & Bax, S. (2005). *Establishing Equivalence in EAP Writing Assessment: Effects of Computerization on Candidate's Cognitive Processing and Test Scores*. Unpublished IELTS report.
- Woods, P. (1992). Symbolic interactionism: theory and method. In M. LeCompte, W. Millory, & J. Preissle, *The Handbook of Qualitative Research in Education* (p. 447-488). London: Academic Press.
- Yu, G. Rea-Dickins, P. & Kiely, R (2007) The cognitive processes of taking IELTS Academic Writing Task 1. *IELTS Research Reports* 11(6). Retrieved 28/06/2015 from http://www.ielts.org/researchers/research/ielts_writing_test.aspx